NEW TECHNOLOGY DEVELOPMENT FOR NEXT-GENERATION SEQUENCING

by

MELISSA A RANDEL

A DISSERTATION

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

June 2017

DISSERTATION APPROVAL PAGE

Student: Melissa A Randel

Title: New Technology Development for Next-Generation Sequencing

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:

Kryn Stankunas	Chairperson
Eric A. Johnson	Advisor
Karen Guillemin	Core Member
John Postlethwait	Core Member
Victoria DeRose	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

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

Degree awarded June 2017

DISSERTATION ABSTRACT

Melissa A Randel

Doctor of Philosophy

Department of Biology

June 2017

Title: New Technology Development for Next-Generation Sequencing

Next-Generation Sequencing (NGS) technologies have been evolving at an unparalleled pace. The ability to generate millions of base pairs of data in a short time and at lower cost than previously has led to a dramatic expansion of technologies within the field. This dissertation discusses the development and validation of new methods for assessing genomic variation, dynamic changes in gene expression, high-accuracy sequencing, and analysis of recombination events.

By reducing the cost of analyzing many samples for genetic divergence by genotyping the same region of the genome in multiple samples, researchers can pursue investigations on a larger scale. Next-RAD (Nextera fragmentation with Restriction-Associated Digestion) allows analysis of a uniform subset of loci between organisms for comparison of populations by genetic differences with reduced burdens of cost and data analysis. This method was applied to the *Anopheles darlingi* mosquito to identify three distinct species that were thought to be a uniform population.

The lowering cost of large-scale sequencing investigations allows for massively parallel analysis of genomic function in a single assay. Regulation of gene expression in response to stress is a complex process which can only be understood by analyzing many pathways in tandem. A novel method is described which quantifies on a genome-wide

scale the expression of millions of randomer tags driven by associated transcriptional enhancers. This method provides novel data in the form of high-resolution analysis of gene regulation.

Aside from generating novel data types, another force behind development of new technologies is to improve data quality. One limitation of NGS is the inherent error rate. PELE-Seq (Paired End Low Error Sequencing) was developed to address this problem, by employing completely overlapping paired-end reads as well as a dual barcoding strategy to eliminate incorrect sequences resulting from final library amplification. This new tool improves data quality dramatically.

Finally, the rapid expansion of tools necessitates the identification of new applications for these technologies. To this end, 10x Genomics Linked-Read sequencing was employed to identify recombination events in multiple species. The haplotype-resolved nature of the data generated from such assays has many promising applications.

This dissertation includes previously published, co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Melissa A Randel

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
University of California, Davis

DEGREES AWARDED:

Doctor of Philosophy, Biology, 2017, University of Oregon
Bachelor of Science, Biotechnology, 2009, University of California Davis

AREAS OF SPECIAL INTEREST:

Genomics
Next-Generation Sequencing
Molecular Biology

PROFESSIONAL EXPERIENCE:

Courtesy Intern, Hui Zong lab, University of Oregon, Jan 2011-Sep 2011

GRANTS, AWARDS, AND HONORS:

Best Poster Presentation, Graduate Student Research Forum, "Functional Genetic Cis-Regulatory Element Identification in Innate Immune Response." University of Oregon, Eugene, OR, 2016

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

NRSA Molecular Biology and Biophysics training grant, University of Oregon, 2011-2014

PUBLICATIONS:

Emerson K. J., Conn J. E., Bergo E. S., **Randel M. A.**, & Sallum M. A. M. (2015). Brazilian Anopheles darlingi root (Diptera: Culicidae) clusters by major biogeographical region. *PLoS One*, 10(7), e0130773.

Kamps-Hughes N., Preston J. L., **Randel M. A.**, & Johnson E. A. (2015). Genome-wide identification of hypoxia-induced enhancer regions. *PeerJ*, 3, e1527.

Preston J. L., Royall A. E., **Randel M. A.**, Sikkink K. L., Phillips P. C., & Johnson E. A. (2016). High-specificity detection of rare alleles with paired-end low error sequencing (PELE-seq). *BMC genomics*, 17(1), 464.

ACKNOWLEDGMENTS

I wish to express sincere thanks to Eric Johnson for providing guidance and a multitude of collaborative projects to work on throughout my graduate studies. Jessica Preston provided guidance and support throughout on data analysis and surviving graduate school. Nick Kamps-Hughes was an excellent rotation mentor and collaborator. Ariel Royall also contributed insight on various projects. Paul Etter was invaluable for his insight on sequencing library preparation and training on the use of several instruments, as well as patiently enduring my numerous questions. I am grateful for the investments of my dissertation committee members Kryn Stankunas, John Postlethwait, Victoria DeRose, and Karen Guillemin. Finally, I would like to acknowledge Maggie Weitzman and Doug Turnbull from the Genomics and Cell Characterization Core Facility (GC3F) for their assistance with multiple projects. My investigations were supported by the META Center for Systems Biology grant (NIH) as well as the NRSA Molecular Biology and Biophysics training grant.

Dedicated to my mother, Monica Randel, who is the strongest woman I know.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Next-Generation Sequencing (NGS) is a rapidly evolving technology platform that is revolutionizing our understanding of genomics. We have only begun to explore potential applications. Over the past decade, NGS technologies have evolved to address some of the technical deficiencies of Sanger sequencing [1], namely high cost of sequencing and error rates [2]. These methods have been applied to human medical research to identify disease-associated polymorphisms and improve treatment methodologies [3] and to rapid and accurate population-level genotyping on a massive scale [4-5]. In this work, we describe the development of several novel NGS technologies for selective sequencing of subsets of loci [6], analysis of dynamic gene regulation by a genome-wide enhancer screen [7], near-elimination of error rates in Illumina sequencing [8], and identification of recombination events by linked-read sequencing.

Next-RAD genotyping

Restriction site Associated DNA Sequencing (RAD-Seq) allows for analysis of a subset of genomic loci between individuals for low-cost analysis of population-level genetic variation [9-11]. Utilizing this method, researchers are able to discover single nucleotide polymorphisms (SNPs) at the same genomic loci in hundreds of individuals with highly reduced sequencing costs compared to whole genome analysis. This core method has been applied to several investigations in a wide variety of species ranging from plants to human samples, and has important implications for conservational biology [12].

In RAD-Seq, the genome of interest is digested with restriction enzymes that have a high-fidelity recognition site. Following digestion, sequencing adapters are ligated to the restriction enzyme cut site such that only fragments adjacent to these cut sites are amplified and sequenced. Polymorphisms adjacent to RAD sites can thus be identified without sequencing the entire genome of interest for the purposes of analyzing population-level variation at low cost.

There have been several methods of modifying RAD-Seq technology to adapt it to different purposes. These include double digest RAD-Seq (ddRAD), which utilizes two restriction enzymes simultaneously to select for sites that contain both recognition sequences [13], quaddRAD, which includes additional multiplexing capabilities as well as PCR duplicate removal [14], and Next-RAD, which was utilized in this investigation to determine population genetic differences in the Anopheles mosquito. Next-RAD genotyping utilizes a Nextera (Illumina, Inc.) reaction to fragment genomic DNA and amplify using modified primers which are complimentary to a selective 8 nucleotide sequence (similar to a RAD site); therefore, only fragments containing this recognition site are amplified. This effectively amplifies consistent genomic loci between samples with the simplicity of a Nextera enzymatic fragmentation reaction [6,12]. This method was developed by Paul Etter and Eric Johnson, and implemented in this investigation with their guidance.

Enhancer activity in stress response

Another aspect of NGS technologies is their ability to dissect genome function in a highly parallel fashion. For functions such as stress response, where the genome is functionalized in real time to respond to environmental change, only a high-throughput

method could encompass all factors of such a complex system. In response to stress, our bodies activate genes via multiple signaling pathways which interact to provide highly sensitive and variable gene expression profiles.

Enhancers are cis-regulatory elements that bind transcription factors to initiate target gene expression. The high-throughput enhancer screen described in this work was applied to characterize response of Drosophila Schneider 2 cells to hypoxia. This method utilizes a library composed of fragmented genomic DNA upstream of a minimal promoter driving expression of nucleotide randomer sequences which can be quantified as a read-out of enhancer activity in their associated genomic fragments [7]. This high-throughput enhancer screen was developed by Nick Kamps-Hughes, with contributions by this author as well as Jessica Preston and Eric Johnson.

High-accuracy sequencing

When dealing with such large datasets, even a low error rate can result in thousands of incorrect nucleotide calls. Polymorphic loci allow for adaptation given selective pressure on the genome, but can consist of very low-frequency alleles within a population which are hard to identify given current NGS technologies. This is highly relevant to evolution as well as cancer biology, as tumors contain many genetic variants at low frequencies that may include drug resistance alleles.

Paired End Low Error Sequencing (PELE-Seq) [8] is a method for investigating the dynamics of ultra-rare alleles which is compatible with standard short read sequencing methodologies. This method encompasses a wet lab protocol and bioinformatics tools to improve NGS data quality by utilizing two strategies: completely

Overlapping Read Pairs (ORPs) to eliminate sequencing errors and a dual barcoding strategy to eliminate PCR errors.

This method was successfully applied to several investigations and was able to identify rare variants at less than 1% of the population with zero false positive SNP calls. By comparison, standard NGS libraries included 30-50% false positive SNP calls. PELE-Seq allows for the detection of ultra-rare alleles that are impossible to reliably identify using traditional NGS methods, even at high quality and coverage. This method was developed by Jessica Preston with contributions by this author as well as Eric Johnson, Ariel Royall, Kristin Sikkink, and Patrick Phillips.

Identification of recombination events using linked reads

Crossover events during the production of gametes allow for the generation of unique offspring with a mix of traits from either parent. This process, called recombination, occurs during meiosis and is an important source of genetic diversity. It is also required for proper assortment of haploid chromosomes into sperm or egg. Recombination rates are variable between species and by sex, thus they must be characterized independently for each organism of interest [15].

The standard method of identifying recombination events and 'hot-spot' loci [16-18] (loci with a higher frequency of recombination events) using current NGS technologies requires sequencing genomes of both parents and their progeny, effectively tripling the cost of producing a genetic map. We utilized a recently developed linked read technology to resolve haplotypes and identify recombination events directly. Haplotype-resolved phase blocks can be constructed using linked reads, which allows for the identification of individual molecules of DNA containing SNPs pertaining to both

haplotypes as indicative of a recombination event. This method was developed and executed by the author in conjunction with Eric Johnson.

CHAPTER II

BRAZILIAN ANOPHELES DARLINGI ROOT CLUSTERS BY MAJOR BIOGEOGRAPHICAL REGION

This work was published on July 14, 2015 in the journal PLoS One by authors Kevin J. Emerson, Jan E. Conn, Eduardo S. Bergo, Melissa A. Randel, and Maria Anice M. Sallum. I performed all wet-lab work for the preparation of sequencing libraries by a novel method, NextRAD, described here, and contributed to authorship.

INTRODUCTION

Anopheles (Nyssorhynchus) darlingi Root is broadly distributed in Central and South America, extending from southeastern Mexico to northern Argentina and from east of the Andes to the Atlantic coast [1]. This species is the most aggressive and effective Neotropical malaria vector, primarily in the Amazon/Solimões River basin. Furthermore, *An. darlingi* is associated with malaria dynamics in forest areas where the natural ecosystems are undergoing intensive ecological changes promoted by deforestation and land use [2, 3].

Anopheles darlingi was described by Root [4] based on morphological characters of the egg, fourth-instar larva, pupa, male and female collected in Caxiribú in the vicinity of Porto das Caixas, Rio de Janeiro state, Brazil. Galvão et al. [5] expanded the geographical distribution of the species to inland São Paulo state, Bahia, and northern Brazil. *Anopheles paulistensis* Galvão, Lane and Corrêa was described as a morphological variant of *An. darlingi* based on differences in the egg, male and female morphology of specimens from Pereira Barreto, inland São Paulo state and Manaus,

Amazonas state [5]. Later, Lane [6] considered that those differences represented phenotypic variations, and *An. paulistensis* was synonymized with *An. darlingi.*

Polymorphisms were also observed in the banding pattern of the X and all four autosome arms of the salivary gland polytene chromosome of representatives of *An. darlingi* populations from three northern localities in the Amazon forest and one southern locality in the domain of Cerrado, inland São Paulo state, and considered to be linked with distinct vectorial capacity [7]. More recently, Malafronte et al. [8] observed intraspecific variability in the rDNA ITS2 sequences that corroborated the northern / southern population polymorphisms in the polytene chromosomes detected by Kreutzer et al. [7]. Furthermore, heterogeneities were also observed in the peak biting behavior [9, 10], in wing morphometric geometry [11], in vectorial capacity [12], and in the genetic structure of southeastern and northern populations using both mtDNA Cytochrome Oxidase I (COI) [13], and microsatellite markers [14]. In contrast, *An. darlingi* has been considered to be a monotypic species based on other data sets [15, 16].

Using specimens spanning almost the entire distribution of *An. darlingi*, COI sequences [17] and microsatellite loci [18] detected deep geographic differentiation that separates Amazonian South America populations from those in Central America, northwestern Colombia and Venezuela. Ancient evolutionary processes were invoked to explain the COI split [17]; in contrast, distance and differences in effective population sizes best explained the level of differentiation detected by microsatellites [18].

Within South American populations, variation in COI resolved two genetic clusters that coincide with two centers of endemism: 1) within the Amazonas/Solimões river basin plus Guyana (north of the Amazon), and 2) within South America (Belém,

Pará), with expansions that occurred during the Pleistocene [17]. Subsequently, it was found that the population growth of *An. darlingi* was not homogeneous [13].

Geographical barriers represented by the rivers Amazonas/Solimões, the Andes, and the coastal mountain ranges in eastern Brazil resulted in at least four subgroups within the South American cluster [13]. It is worthwhile noting that the populations from the lowlands along the Atlantic coast in Rio de Janeiro and Espírito Santo states were markedly distinct from those of central Amazonia, southern and northeast Brazil.

The Atlantic Forest, originally approximately 150 million hectares, is one of the largest tropical rainforests in the Americas. Its extreme latitudinal dimension (about 29 degrees) and an altitudinal span from sea level (Atlantic coast) to ~2800m (Serra do Mar and Serra da Mantiqueira), incorporates tropical and subtropical zones with diverse environmental conditions [19]. The variable landscape, ecology and terrain favor high biological diversity and multiple areas of plant and animal endemism [20, 21]. In this context, Pedro and Sallum [13] demonstrated that populations of *An. darlingi* from the southeastern and inland Atlantic Forest differ substantially, and hypothesized that the major geographic barrier represented by the coastal mountain range limited the dispersal of populations across the Atlantic Forest.

The Neotropical region consists mainly of forest biomes, with some extensive open vegetation biomes along a wide diagonal that comprises the Pampa, Chaco, Cerrado and Caatinga provinces [22]. Gradual development of this open vegetation promoted the separation of one former region into two: 1) northwestern South America and Amazonian forests; and 2) Parana and Atlantic forests [23]. Based on results of a rigorous cladistic biogeographical analysis of 30 plant and animal taxa, Morrone [22] proposed a system of

natural sub-regions and dominions, provinces and districts, which have been categorized into hierarchical levels linked to major tectonic and geological events. At least some of the differentiation observed in *An. darlingi* populations may be attributed to biogeographical events that delineated the Neotropical region. We hypothesize that the development of the open vegetation area comprising the Chacoan dominion, also known as the Chaco, Cerrado and Caatinga biomes, is one of the primary isolating mechanisms that promoted the genetic differentiation of *An. darlingi* population groups (central Amazonia, southern Brazil and southeastern Brazil) proposed by Pedro and Sallum [13].

Herein, we use genotyping by sequencing with nextRAD (nextera-tagmented, Reductively Amplified DNA) markers (Etter et al, paper in preparation) to detect SNPs, which increase marker-resolution approximately three orders of magnitude compared with previous population genetic studies in *An. darlingi* [8, 13–15, 17, 18, 24, 25]. We propose to: 1) assess the level of structure among populations of *An. darlingi* throughout Brazil; 2) address how genetic diversity is distributed between and within the major forest domains of Amazonia and Atlantic Forest compared with Cerrado; 3) examine whether divergence among population subgroups from the Atlantic coast and central Amazonia, southern and northeast Brazil [13], are consistent with the early morphological division proposed between the variant *An. paulistensis* and *An. darlingi*; 4) address the hypothesis that the Amazonian population represents an unknown putative species; and 5) discuss patterns of structure in the context of Neotropical biogeographical regionalization [26].

MATERIALS AND METHODS

Field Mosquito Sampling Strategy

Specimens of *An. darlingi* were chosen from field collections in twelve states in Brazil (Table 2.1) to represent two major subregions proposed by Morrone [2]: 1) Brazilian subregion (AC, AM, AP, MT, PA, RO), and 2) Chacoan subregion (ES, MG, PR, RJ, SP, TO) (Fig 2.1, Table 2.2). Populations from the Chacoan subregion were subdivided into Parana dominion, which includes the Parana Forest province, here named West Atlantic Forest population (MG, PR, and the two more southern SP sampling localities; Figure 2.1) and the Atlantic Forest province, here designated as southeast population (ES, RJ). In addition, sampling from the Chacoan subregion included representatives from the Cerrado province (the northwestern SP sample locality, TO) of the Chacoan dominion. Individuals of the Brazilian subregion were from the South Brazilian dominion (AC, MT, PA, RO) and the Boreal Brazilian dominion (AM, AP) (Figure 2.1), here named Amazonian population.

Mosquitoes were captured either as larvae/pupae or adults. Males and females were collected using Shannon traps. Both adults and immature stages were sampled from multiple habitat types, such as riverside, lakeside, large farm, natural reserve and agricultural settlement, to maximize within region heterogeneity and to reduce the risk of collecting related individuals, particularly in larval habitat.

State	State code	Collection date	Latitude	Longitude	# Individuals
Acre	AC	July 2006	-10.1233	-66.9107	6
Amapá	AP	July 2006	-0.2131	-50.9722	5
Amazonas	AM	Feb 2009	-2.6208	-60.9439	3
Espírito Santo	ES	Oct 2007	-19.0834	-39.8844	4
Minas Gerais	MG	Nov 2006	-20.0021	-49.0795	3
Pará	PA	Oct 2008	-2.7465	-54.227	5
Paraná	PR	May 2007	-24.2715	-54.2906	6
Rio de Janeiro	RJ	May/June 2007	-22.6333	-42.3	6
Rondônia	RO	Jan 2008	-8.7667	-63.9	2
Mato Grosso	MT	May 2007	-9.4108	-59.0228	2
São Paulo*	SP1	Apr 2012	-20.5572	-51.0152	4
São Paulo*	SP2	May 2009	-22.1347	-48.3917	1
São Paulo*	SP3	May 2009	-22.0757	-48.4374	4
Tocantins	TO	Jul 2009	-10.5859	-49.6898	6

Table 2.1. Sampling localities information and their respective geographical coordinates by state in Brazil.

State	State code	Subregion	Dominion	Province	Inferred cluster
Amapá	AP	Brazilian	Boreal Brazilian	Roraima	3
Amazonas	AM	Brazilian	Boreal Brazilian	Imerí	3
Acre	AC	Brazilian	South Brazilian	Rondônia	3
Mato Grosso	RO	Brazilian	South Brazilian	Madeira	3
Rondônia	MT	Brazilian	South Brazilian	Madeira	3
Pará	PA	Brazilian	South Brazilian	Madeira	3
Tocantins	TO	Chacoan	Chacoan	Cerrado	3
São Paulo	SP1	Chacoan	Chacoan	Cerrado	2
Minas Gerais	MG	Chacoan	Parana	Parana Forest	2
Paraná	PR	Chacoan	Parana	Parana Forest	2
São Paulo	SP2, 3	Chacoan	Parana	Parana Forest	2
Espírito Santo	ES	Chacoan	Parana	Atlantic Forest	1
Rio de Janeiro	RJ	Chacoan	Parana	Atlantic Forest	1

Table 2.2. Sampled populations, including the inferred genetic clusters, subdivided into biogeographical subregions, dominions and provinces proposed by Morrone [26].

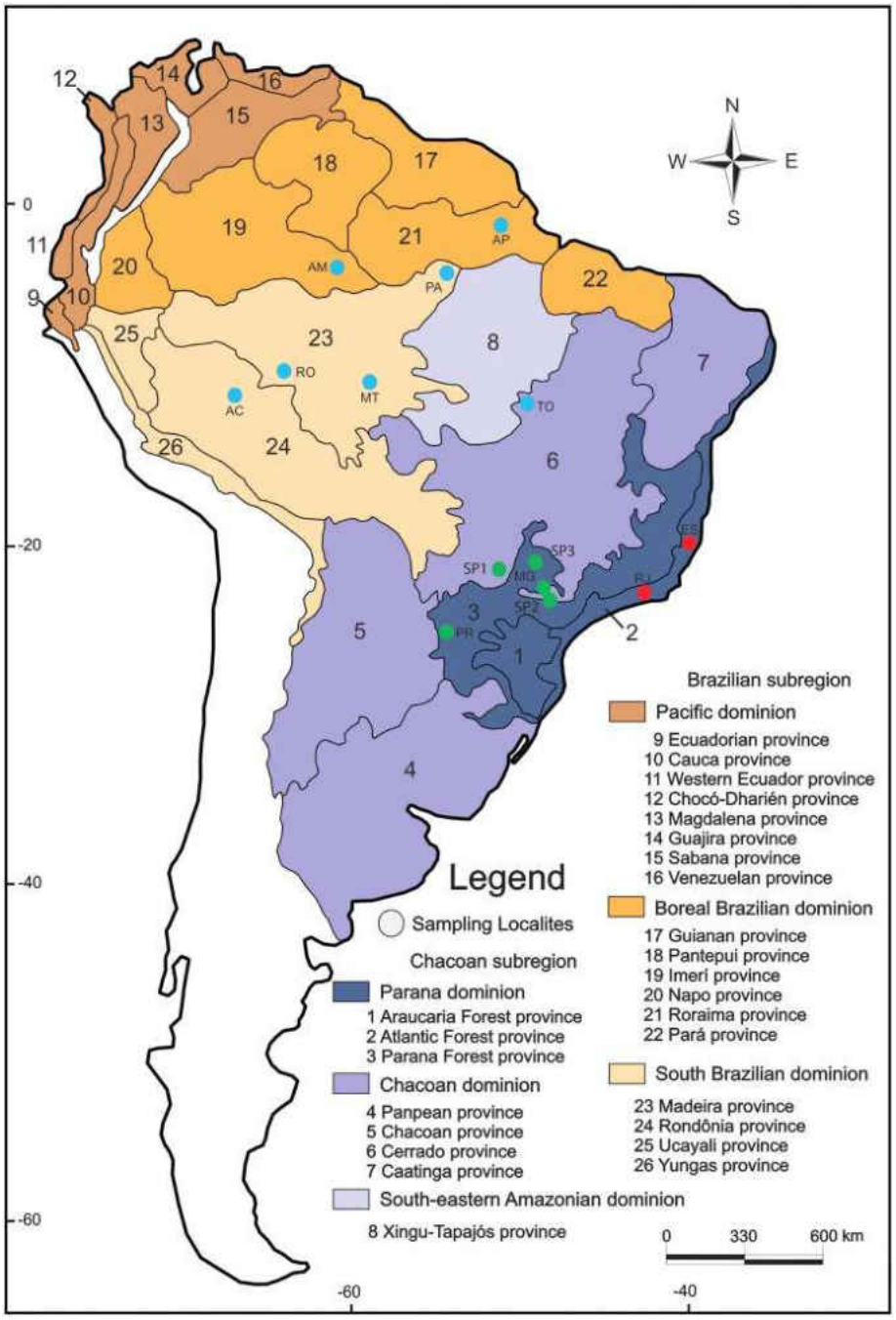

Figure 2.1. Collection sites of *Anopheles darlingi* in relation to biogeographical classification of the Neotropical region proposed by Morrone [26]. Colored (blue, red and green) circles represent the inferred genetic clusters provided by results of STRUCTURE analysis.

All necessary permits were obtained for the described field studies. Collections were made under permanent permit number 16938–1 from Instituto Brasileiro do Meio

Ambiente e dos Recursos Naturais Renováveis (IBAMA) to Maria Anice M. Sallum and E. S. Bergo. Specific permission was not required for these locations as permission to collect was granted under the permanent permit. The collection locations were not privately owned or protected in any way. The field studies did not involve protected or endangered species.

DNA Extraction and Modified Nextera DNA Sample Preparation

Genomic DNA was extracted (Qiagen DNAEasy kit) from 57 individual mosquitoes representing 12 populations (SP1, SP2 and SP3 are a single population; Table 2.1). The DNA was then dried, stored, and later prepared following nextRAD protocols. The nextRAD method uses a selective PCR primer to amplify genomic loci consistently between samples. Genomic DNA (7.5 ng) was first fragmented using a 1/10th Nextera reaction (Illumina, Inc), which also ligates short adapter sequences to the ends of the fragments. Fragmented DNA was then amplified using Phusion Hot Start Flex DNA Polymerase (NEB), with one of the Nextera primers modified to extend 8 nucleotides into the genomic DNA with the selective sequence TGCAGGAG. Thus, only fragments starting with a sequence that can be hybridized by the selective sequence of the primer were efficiently amplified. The following PCR parameters were used: 72°C for 3 minutes, 98°C for 3 minutes, 24 cycles of 98°C for 45 seconds followed by 75°C for 1 minute, then hold at 4°C. The dual-indexed samples were pooled and the resulting library was purified using Agencourt AMPure XP beads at 0.75 X. The purified library was then size selected to 350–500 base pairs. Sequencing was performed in 101-cycles in one lane of an Illumina HiSeq2000 (Genomics Core Facility, University of Oregon).

STACKS and Population Genetic Analyses

Raw Illumina sequences (NCBI SRA Accession numbers SRS950393-SRS950449) were processed with STACKS v1 [27, 28]. Briefly, the raw sequences were quality-filtered using the STACKS program process_radtags. Each of the quality-filtered reads was mapped to the *An. darlingi* genome using bowtie [29]. The reference-genome mapped sequences were then analyzed with STACKS program ref_map.pl. Genotype assignments were corrected using the automated correction module rxstacks. A single SNP position from each RAD locus that had a minimum allele depth of 5 sequences and was scored in at least 50% of individuals within a population was retained and all of these SNP positions used for STRUCTURE analysis [30] for K values between 1 and 8, with 20–40 replicates for each K value. This analysis used a custom script that allows for parallel processing of STRUCTURE analyses (genome.smcm.edu/emersonLab/software). STRUCTURE was run with the admixture model and correlated allele frequencies, and each run used a burnin of 100,000 generations and ran an MCMC chain of 1,000,000 generations. To determine the optimal value of K for our samples, we used the Evanno method [31] implemented in structureHarvester [32]. A complete bash script outlining the parameters used for each component of the STACKS pipeline is provided. Further analysis used a limited SNP dataset that included only those loci (n = 786) that were genotyped in > 75% of individuals in each of the three clusters determined by the full SNP dataset STRUCTURE results. Principle Components Analysis was performed using the R package SNPRelate [33] and AMOVA analysis was performed using Arlequin 3.5 [34].

Due to the possibility of bias introduced in model-based (i.e., STRUCTURE) analyses, particularly due to relatively low numbers of sequences at each locus, we also implemented a Discriminant Analysis of Principal Components (DAPC) [35], implemented in the R package adegenet [36], that does not make any assumptions about the underlying population genetic models. The number of clusters inferred was determined by 100 replicate iterations of K-means clustering using the find.clusters algorithm in adegenet [36].

RESULTS

NextRAD genotyping

An average of 1,625,745 (range: 229,304–5,965,810) 101bp, Illumina reads were aligned to the *An. darlingi* reference genome [37] and resulted in genotype calls at 18,027 (+/- 7,469 SD) loci per individual. Within individuals, 10.83% +/- 0.37 SE loci were heterozygous. Initial filtering of the SNP dataset to include only loci that were genotyped in a majority of individuals from at least one geographical region resulted in a total of 11,533 loci.

Clustering of individuals

There is no evidence of isolation-by-distance among the 12 populations surveyed (Mantel test: $r = 0.02$, $P = 0.36$) that cover a range of 219 to 3,059 km. Therefore we used STRUCTURE [30], Principal Components Analysis, and Discriminant Analysis of Principal Components (DAPC) to further dissect levels of population structure [38].

Based on 11,553 loci, Bayesian clustering analysis via STRUCTURE supports three genetic clusters of *An. darlingi* in Brazil: (1) cluster 1 consists of individuals from

Atlantic Forest province (= southeast) populations (ES and RJ), (2) cluster 2 consists of Parana Forest province, with one Chacoan dominion population (= West Atlantic forest) (PR, SP, MG), and (3) cluster 3 consists of Brazilian dominion, with one Chacoan dominion population (= Amazonian) (AM, AC, AP, MT, PA, RO, TO).

Filtering of the SNP dataset

Once this initial level of population structure was assessed, the genotype dataset was further filtered in order to minimize the possible bias on population genetic inferences due to missing genotype data [39]. The majority of loci genotyped were only scored in one or two of the three genetic clusters (Figure 2.2). Of the 11,533 loci for which genotypes were reliable inferred 1,555 loci were genotyped in individuals from all three clusters and 786 loci were genotyped in > 75% of individuals in each of the three genetic clusters. This filtered dataset of 786 loci was used for downstream analysis.

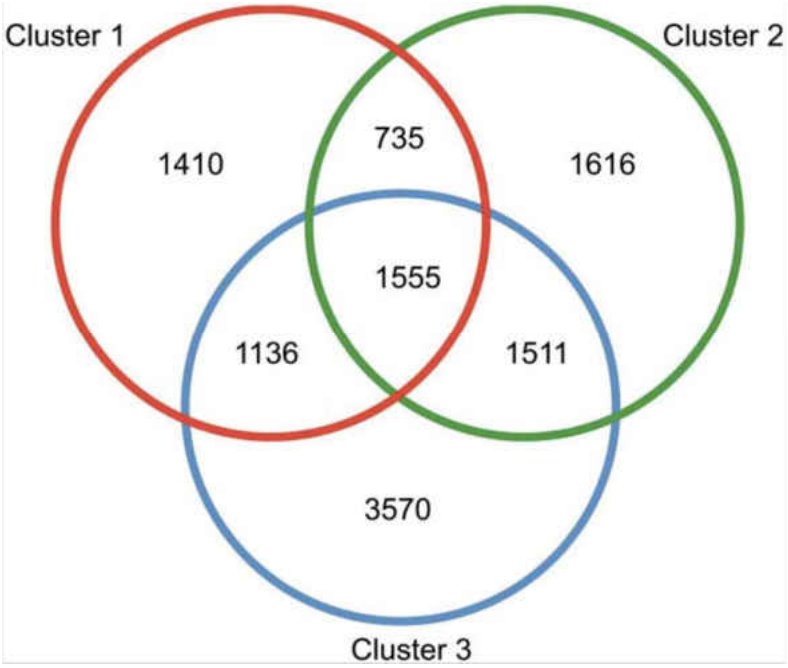

Figure 2.2. Venn diagram showing the number of private and shared genotyped loci of *An. darlingi*, based on loci that were genotyped in at least 50% of individuals from each cluster.

Population genetic inference

STRUCTURE analysis of the filtered SNP dataset discriminated three distinct genetic clusters as outlined below (Figure 2.3B and 2.3C). There were very low levels of allele sharing present, with one individual from cluster 2 showing mixing with cluster 1, and two individuals from cluster 3 showing mixing with cluster 2.

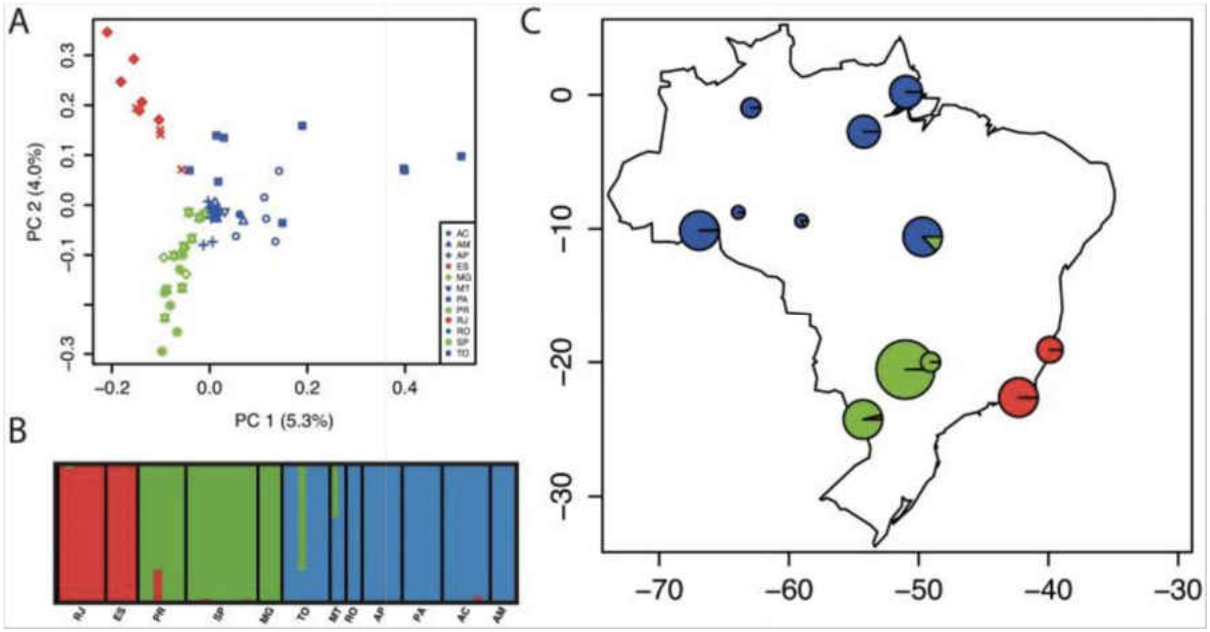

Figure 2.3. Results of Principal Components Analysis (PCA) and STRUCTURE analysis of *Anopheles darlingi* populations using the filtered SNP dataset (786 loci).

Principal Components Analysis (PCA) showed clear partitioning of the populations in the first two principal components (Figure 2.3A). The first principal component (PCA1 5.3%) clearly discriminated the Amazonian (cluster 3) and non-Amazonian (clusters 1 and 2) populations, and the second principal component (PCA2: 4.0%) discriminated the non-Amazonian populations. Coefficients of inbreeding were all not significantly different than zero (Table 2.3).

Cluster	Description	N	Average number of individuals per locus	Genome-wide Pi ± SE	Fis
1	Southeast	10	7.6	0.0766 ± 0.0477	-0.0498[NS]
2	West Atlantic	18	14.4	0.0923 ± 0.0611	-0.0396[NS]
3	Amazon	29	22.4	0.1296 ± 0.0765	-0.0134[NS]

[NS] Not Significant

Table 2.3. Summary statistics for the three inferred clusters of *Anopheles darlingi*.

In the DAPC analysis, there was no clear 'best' value for the number of clusters, with the Bayesian Information Criterion (BIC) value for one, two, or three clusters, being very similar (Fig 4A). Therefore we consider both the case where there are 2 (Fig 4B) and 3 (Fig 4C) clusters. If genotypes are partitioned in to two distinct clusters, there is a clear delineation of the Atlantic Forest populations (cluster 1 above) from the Amazon and Parana Forest populations (clusters 2 and 3 above) (Fig 4B). If we partition our genotypes in to three distinct clusters, the clusters are identical to those from the STRUCTURE analysis. We assessed the robustness of these results by performing one hundred replicate analyses using the algorithm find.clusters (from adegenet [36]) for each of the above clustering schemes and individuals were always placed in to the same clusters.

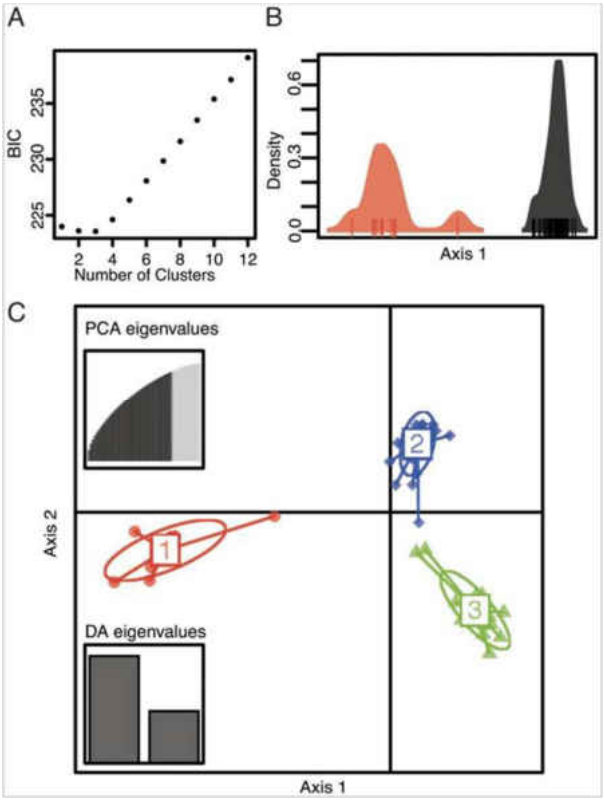

Figure 2.4. Summary of the discriminant analysis of principal components (DAPC). There were significant levels of pairwise genetic divergence among the three clusters (AMOVA, overall Fst = 0.20, P < 0.001) with the highest genome-wide divergence

between the southeast and West Atlantic populations: southeast population—West Atlantic population (Cluster 1 –Cluster 2; Fst = 0.11, P < 0.01), southeast population—Amazon population (Cluster 1 –Cluster 3; Fst = 0.06, P < 0.01), and West Atlantic population—Amazon population (Cluster 2 –Cluster 3; Fst = 0.06, P < 0.01). There was also significant level of genetic divergence between the multiple Amazonian populations as compared with the non-Amazonian populations (Fst = 0.05, P < 0.01).

DISCUSSION

Reduced representation genomic library methods, including nextRAD, suffer from sampling biases as there are usually large numbers of loci that are genotyped in only one or a few individuals [40]. Simulations have shown that datasets that are filtered to minimize the amount of missing data are more likely to accurately reflect population genetic inferences [39]. Under such filtering schemes, loci that are more highly divergent among samples tend to be excluded from the filtered datasets and thus any derived estimates of divergence are likely underestimates of true divergence values. In the data presented here, of the ~11,000 loci that were reliably genotyped in more than 50% of individuals in at least one cluster, only 768 loci were genotyped in more than 75% of individuals in all clusters. The smaller, filtered dataset was used for the majority of analyses to minimize the impact of bias due to the genotype sampling.

Support for geographical differentiation in *An. darlingi* depends on the markers scored and the locations sampled, similar to results in other mosquitoes (e.g. [41, 42]). For single-locus COI gene sequences, Mirabello & Conn [17], studying sampling locations spanning distances from 2–4,870 km, detected the highest levels of genetic differentiation between Central America and northern Amazonia, even though specimens from São Paulo and Mato Grosso states, both south of the Amazon River, were included in the analysis. Within the Brazilian Amazon [14, 25] and between Central and South

America [18], microsatellite markers detected highly significant geographic differentiation. Pedro and Sallum [13], by including individuals representing the Atlantic Forest and Parana Forest provinces of the Parana dominion, Chacoan subregion, found strong evidence of population splits that are primarily coincident with the Chacoan and Brazilian subregions proposed by Morrone [26]. Even though microgeographic differentiation was not detected between neighboring Colombian states [43], Angêlla et al. [24] identified two genetically distinct sub-populations adapted to different seasonal and climatic conditions in localities along the Madeira River, Rondônia state, Brazil. Taken together these studies imply that Neotropical landscape barriers are primary drivers of divergence in *An. darlingi* at regional and continental scales, and that distance and environmental conditions contribute to differentiation at a local scale.

Several approaches were employed in the present study to address genomic variation among *An. darlingi* populations and to test whether clusters are consistent with well-separated species. Analyses of the genome-wide data showed that individuals group into three genotypic clusters. Cluster 1 (red) comprises populations from the Atlantic Forest province (ES, RJ) of the Parana dominion, representing *An. darlingi*. Cluster 2 (green) includes representatives from localities within the Parana Forest province of the Parana dominion (SP, MG, PR) with one Cerrado province population (Chacoan dominion). Cluster 3 (blue) incorporates the Boreal Brazilian and South Brazilian dominion populations (with one Cerrado province population) (Figure 2.1). Thus, the Cerrado province population is split between clusters 2 and 3. There is significant level of divergence between the Boreal Brazilian and South Brazilian dominion populations. (Amazonian populations) (Cluster 3) and the non-Amazonian populations (Clusters 1 and

2), but this divergence is only 50% of that seen between Clusters 1 and 2. Based on these findings, on low admixture between Clusters 1 and 2 (Figure 2.2), and on previous data demonstrating that a physical barrier, e.g., the Serra do Mar on the Atlantic coast, restricts gene flow between *An. darlingi* populations from the Atlantic Forest province and the remaining populations from the Chacoan and Brazilian subregions [13], we propose that Cluster 2 populations represent putative *An. paulistensis*. Within the western Atlantic forest, there is evidence from studies using multiple markers that the coastal mountain range limits dispersal in the bromeliad malaria vector complex *Anopheles (Kerteszia) cruzii*, such that different putative species have evolved [44, 45]. This finding lends support to our hypothesis of possible species-level differentiation between Clusters 1 (putative *An. darlingi*) and 2 (putative *An. paulistensis*).

Cluster 3 populations represent the Boreal Brazilian dominion (AM, AP) and South Brazilian dominion (AC, MT, PA, RO) both within the Brazilian subregion; in addition, this cluster includes individuals from the Cerrado province (TO) of the Chacoan dominion. There is a low level of allele sharing between clusters 2 and 3. One of these individuals is from Cerrado province (TO) population and the other sample is from Madeira province (MT) (Fig 2.2). The shared polymorphism of a second individual between Cerrado province (TO—cluster 3) and Parana Forest province (cluster 2) suggests that the former is a transition zone, with some attributes of both Amazon and West Atlantic Forest. A similar occurrence was observed in the population from Paraná province in the West Atlantic Forest (cluster 2), with one individual from PR sharing polymorphisms with the southeast cluster 1 (RJ, ES).

If our inference for *An. darlingi*, based on Morrone [22, 26] of possible speciation level divergence between Brazilian (cluster 3) and Chacoan subregions (clusters 1 plus 2), and between Atlantic Forest (cluster 1) and Parana Forest (cluster 2) provinces is accurate, other Neotropical organisms with similar distributions may be expected to show similar biogeographic or phylogeographic patterns. In fact, Costa [46], using data from the mitochondrial cytochrome b gene, observed that small forest-dwelling mammals distributed between and within the major forest domains of the Amazonia and Atlantic Forests and the intervening interior forest of Brazil diverged significantly. Between sister taxa of Neotropical orchard bees, Silva et al. [47] found that climatic oscillations that further separated these two large forest biomes promoted parapatric speciation, in which many species had their continuous distribution split, giving rise to different but related species. In the pantropical tree genus Manikara, the divergence between Atlantic coastal forest and Amazonian clades coincided with the formation of drier Cerrado and Caatinga habitats between them [48]. A clade of the frog Hypsiboas albopunctatus from the central Cerrado was found to have diverged from a southeastern clade (Brazilian Atlantic Forest) during the mid-Pleistocene [49]. Soil microbial acidobacteria 16S rRNA sequences are highly differentiated between Cerrado province (of Chacoan dominion) and Atlantic Forest (of Parana dominion), correlated with the distinctive soil and vegetation in each biome [50].

In addition, Nihei and Carvalho [51] defended the hypothesis that the vast Amazon region is not a biogeographical unit, but it is divided into southeastern and northwestern portions. The southeastern portion is closely related to the Chacoan and Parana dominions. These dominion relationships were inferred based on biogeographical

patterns obtained for species of the genus Polietina (Diptera: Muscidae) from the Neotropical region. The fact that the *An. darlingi* population from Tocantins state (Cerrado province, Chacoan dominion) clustered with populations from the South Brazilian dominion may be a consequence of phylogenetic and biogeographical patterns that promoted the division of the forest biomes of the Neotropical region into the main components postulated by [52]. Consequently, two *An. darlingi* population of the Cerrado province (Chacoan dominion) did not cluster together but split into two clusters representative of the Brazilian dominion (cluster 3) and Parana plus Chacoan dominions (cluster 2). Alternatively, our results may be a consequence of sampling strategy with only two populations from the Chacoan dominion, which did not allow a clear separation among distinct biogeographical components postulated by Morrone [2, 6].

It is noteworthy that *An. darlingi* was described by Root [4] using specimens from a locality in Rio de Janeiro state (RJ) situated within the Atlantic Forest province (Figure 2.1), which clustered with representatives of ES, from the same province. In contrast, the MG, SP and PR populations from the Parana Forest province (with one Cerrado province population—SP) clustered separately. We hypothesize that the Parana Forest province cluster may represent the putative *An. paulistensis*, described by Galvão et al [5] from samples captured in Pereira Barreto, formerly Lussanvira municipality, in the West Atlantic Forest within the Parana Forest province. This species was synonymized with *An. darlingi* by Lane [6]; here we propose that *An. paulistensis* may be a valid putative species of the subgenus Nyssorhynchus. The genetic divergence between clusters 1 and 2 and the fact that cluster 3 is equally divergent from the other two clusters could also indicate that heterogeneous divergence among populations of *An. darlingi* was caused by

ecological selection pressures and historical biogeographical processes that may have allowed the contact and separation among distinct populations during the historical events that had led to major Brazilian biome formation.

Several recent studies have led to the discovery of heterogeneous divergence across anopheline genomes under eco-environmental selection pressure [53–55]. Such investigations have provided details of population differentiation that contribute to a more precise understanding of mechanisms of divergence and speciation of particular interest to vector biology. This is amply demonstrated by critical evidence that the M (*An. coluzzii*) and S (*An. gambiae*) forms, recently described as valid species, continue to differentiate [56]. Further study into the genomic patterns of differentiation in *An. darlingi* may shed light on the mechanisms underlying its significant vectorial capacity in the Neotropics, and also help to clarify the vector status of the species in areas outside and inside the Amazon River basin.

BRIDGE TO CHAPTER III

The method described here, Next-RAD genotyping, is a powerful tool for analyzing genetic divergence at a population level. As a novel technology, it allows for reduced sequencing costs by analyzing a subset of loci. Within the field of NGS technologies, this allows for more efficient use of sequencing resources. The continuous development of NGS technologies is important for reducing costs to allow for ever more large-scale investigations, as well as creating novel data types to help us better understand genome function. The next technology discussed deals with such a novel data type: a genome-wide readout of enhancer activity in response to hypoxia.

CHAPTER III

GENOME-WIDE IDENTIFICATION OF HYPOXIA-INDUCED ENHANCER REGIONS

This work was published on December 21, 2015 in the journal PeerJ by authors Nick Kamps-Hughes, Jessica L. Preston, Melissa A. Randel, and Eric A. Johnson. I contributed to all parts of development of the method, including generation of the reporter library, transfection, exposure to hypoxia, and sequencing library preparation.

INTRODUCTION

Gene expression is differently regulated in different cell types and in response to changes to environmental conditions. This regulation is achieved in part by the activity of enhancers [1-5], specific DNA sequences that bind transcription factors to control the rate of transcription initiated at nearby promoters. Even for relatively simple processes, such as the acute response to changes in oxygen availability, the identification and characterization of the enhancers used to shift the network of gene expression to a new mode remains limited.

The transcription factor hypoxia-inducible factor-1 (HIF-1) is directly inhibited by the presence of cellular oxygen via protein degradation of the HIF-1α subunit6. Once stabilized, HIF-1α moves to the nucleus and up-regulates the transcription of target genes. Although HIF-1 remains a central regulator in models of how cells respond after experiencing low oxygen [7-8], more recently other transcription factors have been implicated in the hypoxic response in a complex network of regulatory events. For example, the immunity response transcription factor NF-KB is also activated by hypoxia

and regulates the transcription of HIF-19 [10], while HIF-1 appears to play a reciprocal role in the regulation of NF-kB targets [11]. Likewise, HIF-1 sensitizes the heat shock response by directly regulating heat shock factor (HSF) transcription during hypoxia. Thus, the broader picture that has emerged is that the stress response transcription factor pathways are not isolated regulatory units but rather cooperate and co-opt each other to modify the cell's functions in a complex manner.

High-throughput sequencing tools have become widespread in gene expression studies [12-14]. For example, RNA-Seq has become a powerful tool for analyzing differential gene expression by quantifying the RNA abundance of the transcriptome. However, RNA-Seq does not provide empirical information about the regulatory events leading to a change in transcript abundance. ChIP-Seq provides information about where transcription factors bind to the genome, but binding events do not always result in an active enhancer or change in the rate of transcription. Other sequencing methods assay open chromatin conformations (DNAse-Seq, FAIRE) as a reliable proxy for enhancers. However, until recently the typical functional assay for enhancers was to clone the putative regulator upstream of a reporter gene driven by a minimal promoter.
Several next-generation sequencing-based methods have been used to dissect the function of individual nucleotides within previously known enhancers [15-18] as well as scan genomic sequence for enhancer activity [19]. Here we use a novel variation on these high- throughput enhancer screening methods to identify regions of the Drosophila genome with increased activity under hypoxia. Our technique combines the sheared genomic fragments to be assayed for activity with a UTR randomer tag system for highly

multiplexed tracking of transcriptional activity. The construct library is modularly synthesized in vitro making the relative placement of construct elements easily mutable.

The work presented here is the first implementation of a massively parallel reporter assay to study cis-regulatory activity during an environmental stress response. A library of 4,599,881 random 400-500bp fragments spanning the Drosophila melanogaster genome was used to identify 31 hypoxic enhancer regions. The regions coincide with genes up- regulated under hypoxia and with binding site motifs from multiple transcription factors involved in the hypoxic response. This work provides mechanistic details of the hypoxic response by empirically identifying regulatory regions that drive hypoxic transcription, linking them to target genes from RNA-Seq differential expression data, and identifying trans-acting factors in silico. This genome-wide scan demonstrates the complexity of the hypoxic response, which involves multiple regulators acting in concert to control the expression of a wide variety of targets.

MATERIALS AND METHODS

All DNA sequencing was performed on the Illumina HiSeq. All PCR reactions contained a final concentration of 400nM of each primer and used Phusion Polymerase in 1X HF buffer.

Library synthesis

The linear reporter library used to assay enhancer activity was constructed entirely in vitro (Figure 3.1A). The sequence space being assayed for enhancer activity, in this case the Drosophila melanogaster genome, was sonically sheared to generate random enhancer-sized fragments. Adapter ligation and 5' PCR addition were used to

add the Illumina first-end sequence upstream of the sheared DNA and part of the minimal promoter downstream. 5' PCR additions are used to add minimal promoter elements, an intron to stabilize mRNAs20, the 20N randomer tag, and Illumina paired-end sequence upstream of an arbitrary ORF, in this case GFP. The synthetic minimal promoter used was designed to contain several core motifs and has been shown to function with a wide range of enhancers [21]. The two fragments are then ligated together to create the final construct library pictured in Figure 3.1A. The reporter library was diluted to a target of 10,000,000 molecules and regenerated by PCR so that the library could be adequately characterized by paired-end sequencing. An aliquot of the reporter library is used for paired-end sequencing to match randomer tags located in the 5' UTR to the non-transcribed genomic region driving their expression. The library is then transfected into cells for massively parallel enhancer assay (Figure 3.1B).

Drosophila melanogaster strain Oregon-R genomic DNA was sonically sheared using the BioRuptor. 400-500bp fragments were isolated by gel electrophoresis then end-repaired using Blunt Enzyme mix (NEB) and 3' adenylated using Klenow exo- (NEB). This sample was then ligated to an asymmetric adapter with T-overhang composed of annealed oligonucleotides Genomic-Adapter-1 and Genomic-Adapter-2. The ligation product was gel-purified and used as PCR template with primers Illumina P5 and Genomic-R to create a library of molecules containing a random 400-500 bp stretch of Drosophila melanogaster genomic sequence between the Illumina end one sequence and the beginning of a synthetic promoter. Separately, The GFP coding sequence followed by the SV40 terminator was PCR amplified from plasmid pGreen-H-Pelican with primers GFP-F and SV40-R. This product was then used as template for a PCR reaction using

primers SV40-R and Marker-1-F. This product was then used as template for a PCR reaction using primers SV40-R and Marker-2-F. This product was then used as template for a PCR reaction using primers SV40-R and Marker-3-F to create a library of molecules containing a GFP sequence downstream of a minimal promoter with randomer tag and Illumina paired-end sequences.

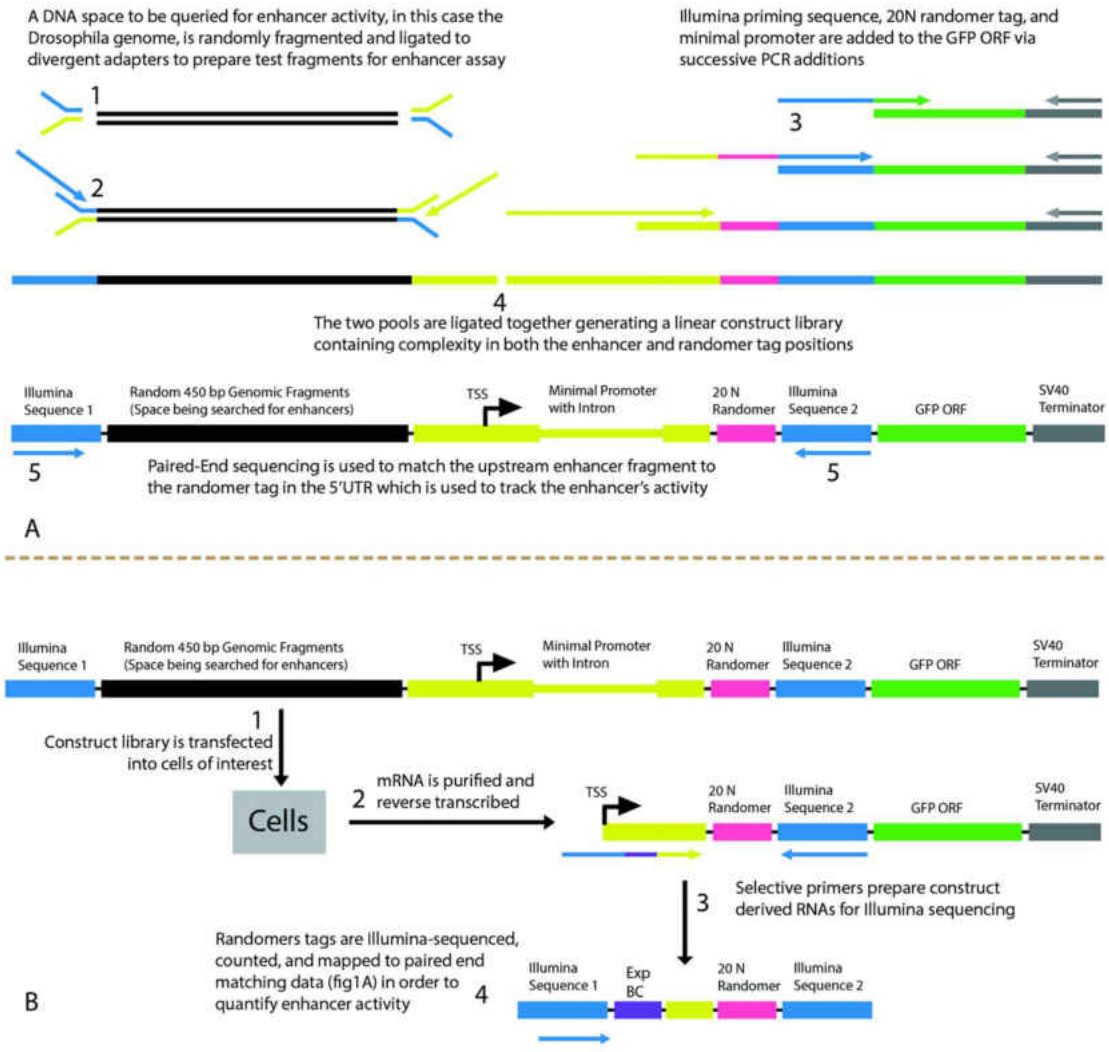

Figure 3.1. Enhancer library synthesis and assay. (A) DNA of interest is fragmented (step 1) and ligated to divergent adapters (step 2) leaving potential enhancer fragments with Illumina sequence on one side and the beginning of the synthetic minimal promoter on the other. The GFP gene is used as a template for a series of 5' PCR additions in order to add Illumina sequence, 20N randomer tag, and the majority of the minimal promoter and intron (step 3). The two sides are ligated together to create a linear construct with complexity in the enhancer region upstream of the transcription start site as well as

complexity in the randomer tag region in the 5' UTR (step 4). The sample is submitted to paired-end sequencing in order to match the potential enhancer region to the randomer tag in the 5' UTR that is used to report its activity. (B) The enhancer library is transfected into cells (step 1) and total RNA is purified and reverse transcribed to create cDNA (step 2). The cDNA is used as template for a PCR reaction (step 3) with a reverse primer complimentary to the Illumina end 2 sequence present in the construct and a forward primer complimentary to the stretch of the minimal promoter upstream of the randomer tag. The forward primer adds Illumina end 1 seqeunce and an experimental barcode for multiplexing. This amplicon is ready to be loaded onto the Illumina flow cell for single-end sequencing of randomer tags (step 4) in order to quantify enhancer activity.

The genomic sequence-containing library and minimal promoter library were then 3' adenylated and 3' thymidylated respectively with Klenow exo- then ligated together. The heterodimer (1819-1919bp) was gel-purified and subsequently selected for proper orientation by PCR with primers SV40-R and Illumina P5. To reduce library complexity to a scale that was tractable by paired-end sequencing, DNA was quantified using the Qubit system (Invitrogen) and serially diluted to produce an estimated 10,000,000 molecules that were used as template to regenerate the library by PCR with primers SV40-R and Illumina P5. An aliquot of this library was used as template for a PCR reaction with primers Illumina-P7 and Illumina-P5 to generate a paired-end Illumina-sequencing library such that the first-end sequence contained the beginning of the genomic region and the paired-end sequence contained the corresponding randomer tag (Figure 3.1A). Aliquots were also used to generate transfectable quantities of the full-length reporter library by PCR amplification of the entire fragment using primers SV40-R and Illumina-P5.

Transfection, RNA extraction, and randomer tag sequencing

Six 5mL flasks were plated to 80% confluency with S2 cells and transfected with Fugene HD and 2.6ug reporter library DNA at a 3:1 ratio. The following day three plates were placed under hypoxia (99.5% N2 and 0.5% O2) for five hours and thirty minutes

and three were left in atmospheric conditions. Total RNA from both conditions was extracted using Trizol and treated with DNAse Turbo (Ambion). RNA was converted to cDNA with SuperScipt III first strand synthesis kit (Invitrogen) using oligo dT20 primers. cDNA was used as template for PCR with primers flanking the randomer tag to create an amplicon ready for Illumina sequencing. All PCR reactions used Illumina-P7 reverse primer and the following barcoded forward primers to allow multiplexing: RNA-BC-1 for hypoxic sample 1, RNA-BC-2 for hypoxic sample 2, RNA-BC-3 for hypoxic sample 3, RNA-BC-4 for normoxic sample 1, RNA-BC-5 for normoxic sample 5, RNA-BC-6 for normoxic sample 6. The resulting 178-bp amplicons were combined and sequenced on the Illumina Hiseq.

RNA-Seq

RNA from the same experiments used to quantify enhancer activity was used for RNA-Seq. mRNA was purified using Dynabeads (Invitrogen) from 10ug of total RNA and chemically fragmented using Ambion Fragmentation Reagent. cDNA libraries were made with SuperScipt III first strand synthesis kit using random hexamer primers followed by second-strand synthesis with DNA Pol I (NEB). The double stranded DNA was end-repaired using NEB Quick Blunting Kit and 3' adenylated using Klenow exo-. The samples were ligated to divergent Illumina adapters with in-line barcodes (Hypoxic GGTTC, Normoxic CTTCC) and PCR amplified with Illumina primers. 300-450 bp fragments were gel-purified and sequenced on the Illumina HiSeq (hypoxic condition: Accession SRX467593, normoxic condition: Accession SRX467591). 6,855,528 reads from each sample were aligned to the Drosophila melanogaster transcriptome (Flybase, r5.22) using TopHat22. The bam outputs were analyzed by cufflinks and the resulting

transcripts.gtf files were compared using cuffdiff to identify differentially expressed genes. Some ncRNAs were also analyzed for differential expression. As they are not present in the transcriptome build, RNA-Seq reads were aligned to each ncRNA using Bowtie223 and their expression level is reported by normalized number of aligned reads in each condition.

Computational enhancer activity analysis pipeline

Paired-end fastq files (Accession SRX468157) linking genomic regions in the first-end read to randomer tags in the paired-end read were parsed to a fasta file with the randomer tag as the sequence name and the genomic sequence as the sequence. This file containing 32,061,029 sequences was aligned to the Drosophila melanogaster genome (NCBI build 5.3) using Bowtie223. Reads were processed into a match-list linking randomer tags to the genomic coordinates of their corresponding test sequence.

Randomer tags from hypoxic and normoxic RNA amplicon sequencing were extracted from fastq files (Accessions SRX468694, SRX468097) and experimental replicates were separated by barcode. 18,261,667 randomer tags from hypoxic sample 1, 14,226,458 from hypoxic sample 2, 14,697,154 from hypoxic sample 3, 14,406,854 from normoxic sample 1, 14,988,132 from normoxic sample 2, and 11,516,478 from normoxic sample 3 were referenced to the paired-end match list to generate genome-wide enhancer activity tables by 100bp bins. The genomic fragments ranged from 400-500bp so the bin corresponding to the alignment as well as the four downstream bins were credited 1 count. In the cases where randomer tags matched multiple genomic fragments, bins were credited a fraction of a count based on the likelihood of that linkage in the paired-end match data. This created a genome-wide count table of enhancer activity in each

replicate. The count table was then analyzed in R for differential activity between hypoxic and normoxic replicates using a negative binomial test in the DESeq24 package.

The bins were filtered by overall count (θ=0.5) and the test was run with default variance estimation. This generated a p-value and a p-value adjusted for multiple hypothesis testing (Benjamini-Hochberg procedure) for each 100bp bin. Hypoxic enhancer regions were defined at bins up-regulated under hypoxia with adjusted p-value < 0.1 (p-value < 1.55 e-05) and extend to include adjacent bins with p-value < 0.05.

Enhancer sequence motif analysis

Identified enhancer regions were searched for stress transcription factor binding sites using the BoBro BBS motif-scanning algorithm [25] with position weight matrices from the JASPAR database [26]. This algorithm was used to identify binding site positions and calculate a global p-value of enrichment for HIF-1 (JASPAR ID: MA0259.1), FOXO (MA0480.1), HSF (MA0486.1) and NF-kB (MA0105.3) binding sites in enhancer sequences compared to the Drosophila melanogaster genome background.

RESULTS

Discovered hypoxic enhancers

Transcriptional activity from 4,599,881 fragments that were 400-500bp in size, spanning the Drosophila melanogaster genome at 17.39X coverage, was analyzed by 100bp bins and 31 significant hypoxic enhancer regions (q-value < 0.1, p-value < 1.55 e-05) were identified (Table 3.1). These enhancer regions range in size from 100 to 800bp and confer 2 to 18-fold changes in expression under hypoxia. The discovered enhancers

are found throughout the genome and are located proximally to genes up-regulated under hypoxia in our RNA-Seq experiments. The ten most strongly up-regulated genes all contain a discovered enhancer within 20kb. 16 of 31 discovered enhancers are located within 20kb of one of the 90 up-regulated genes. The probability of this positional overlap occurring by chance is 1.43 e-14 using an exact binomial test, supporting that the discovered enhancers are linked to endogenous gene expression and implicating their likely targets. 4 additional enhancers are proximal to genes previously observed to be up-regulated under hypoxia in Drosophila [27].

Enhancer Locus	P-value	Adjusted P-value	Fold Change	Hyp. Gene(s) Within 20Kb	Relative Position to Hyp. Gene(s)	Stress TF Binding Sites
3R:8303000..8303500	7.79 e-22	4.63 e-16	5.08	Hsp70B genes	Intergenic	Hsf, Hif-1, Foxo
3L:6256700..6257200	1.83 e-16	2.72 e-11	5.95	impl3	Upstream	NF-kB
3R:8331100..8331800	1.59 e-16	2.72 e-11	4.49	Hsp70Bb	Promoter Proximal	Hsf, Hif-1, Foxo
3R:8293200..8293900	2.96 e-16	3.51 e-11	3.83	Hsp70Ba	Promoter Proximal	Hsf, Hif-1, Foxo
3R:8334400..8335000	1.18 e-15	1.01 e-10	4.45	Hsp70Bc	Promoter Proximal	Hsf, Hif-1, Foxo
2L:8001300..8001800	2.64 e-15	1.74 e-10	6.44	Wwox	Intronic	Hif-1
3R:8327800..8328500	8.89 e-13	2.40 e-08	3.70	Hsp70Bbb	Promoter Proximal	Hsf, Hif-1, Foxo
2L:20082900..20083500	1.08 e-12	2.79 e-08	6.35	Fok	Intronic	Foxo, Hif-1
3L:8685300..8685800	1.07 e-10	2.18 e-06	3.79	Hairy	Downstream	Hsf, Hif-1, Foxo
3L:7797800..7798600	1.77 e-10	3.38 e-06	3.07	CG32369	Intronic	Hif-1
3L:9385200..9385800	2.14 e-09	3.62 e-05	3.71	Hsp22,23,26,27	Neighboring Intron	Not Detected
X:17071000..17071300	8.77 e-09	1.24 e-04	4.99	Not Detected	Not Detected	Not Detected
X:9767000..9767500	1.27 e-08	1.76 e-04	3.65	CG32695*	ORF	Not Detected
2L:2887100..2887600	1.32 e-08	1.79 e-04	5.82	Not Detected	Not Detected	Hif-1
3L:11234100..11234900	6.03 e-07	6.63 e-03	2.68	Scylla	Upstream	Foxo
3L:3892900..3893100	1.55 e-06	1.59 e-02	2.75	Not Detected	Not Detected	Hif-1, NF-kB
2L:5986900..5987500	1.82 e-06	1.81 e-02	2.16	ifc*	Intronic	Foxo
3L:9448800..9448900	2.09 e-06	2.03 e-02	5.39	MTF-1*	Neighboring Intron	NF-kB, Hif-1
3R:6800900..6801600	2.22 e-06	2.09 e-02	13.82	Not Detected	Not Detected	Hif-1
3L:11522800..11523300	2.66 e-06	2.35 e-02	3.04	Not Detected	Not Detected	NF-kB
3R:4181100..4181600	2.66 e-06	2.35 e-02	3.87	Atg13	Downstream	Foxo, Hif-1
3R:7781900..7782700	2.69 e-06	2.35 e-02	4.96	Hsp70Aa	Promoter Proximal	Hsf
3R:7783900..7784500	2.75 e-06	2.37 e-02	4.18	Hsp70Ab	Promoter Proximal	Hsf
3R:21433600..21434000	3.30 e-06	2.72 e-02	9.03	Not Detected	Not Detected.	Not Detected
X:16559200..16559700	4.13 e-06	3.23 e-02	6.56	Not Detected	Not Detected	Foxo
3R:2902300..2902600	6.21 e-06	4.63 e-02	2.95	Not Detected	Not Detected	Not Detected
2R:12896000..12896500	6.88 e-06	5.05 e-02	3.02	Not Detected	Not Detected	Foxo
X:17388000..17388500	8.24 e-06	5.75 e-02	6.80	Not Detected	Not Detected.	Hif-1
3R:14892300..14892800	9.76 e-06	6.44 e-02	18.01	Not Detected	Not Detected	Hif-1
3R:27050000..27050500	1.52 e-05	9.40 e-02	2.78	CG12054*	Intronic	Hif-1
3R:25921500..25922100	1.54 e-05	9.44 e-02	2.46	Hif-1	Intronic	NF-kB, Hif-1

Table 3.1. Properties of discovered hypoxic enhancers. Genes up-regulated under hypoxia are from RNAseq experiments from the same RNA pools used to quantify enhancer activity unless denoted by an asterisk in which case they were observed to be up- regulated under hypoxia in Drosophila by Li et al. [27].

Location of hypoxic enhancers

Of the 20 hypoxic enhancer regions proximal (within 20kb) to hypoxic up-regulated genes, 6 fall in the promoter region of the putative target gene (Figure 3.2, Table 3.1). All six of these are the homologous Hsp70B enhancers.

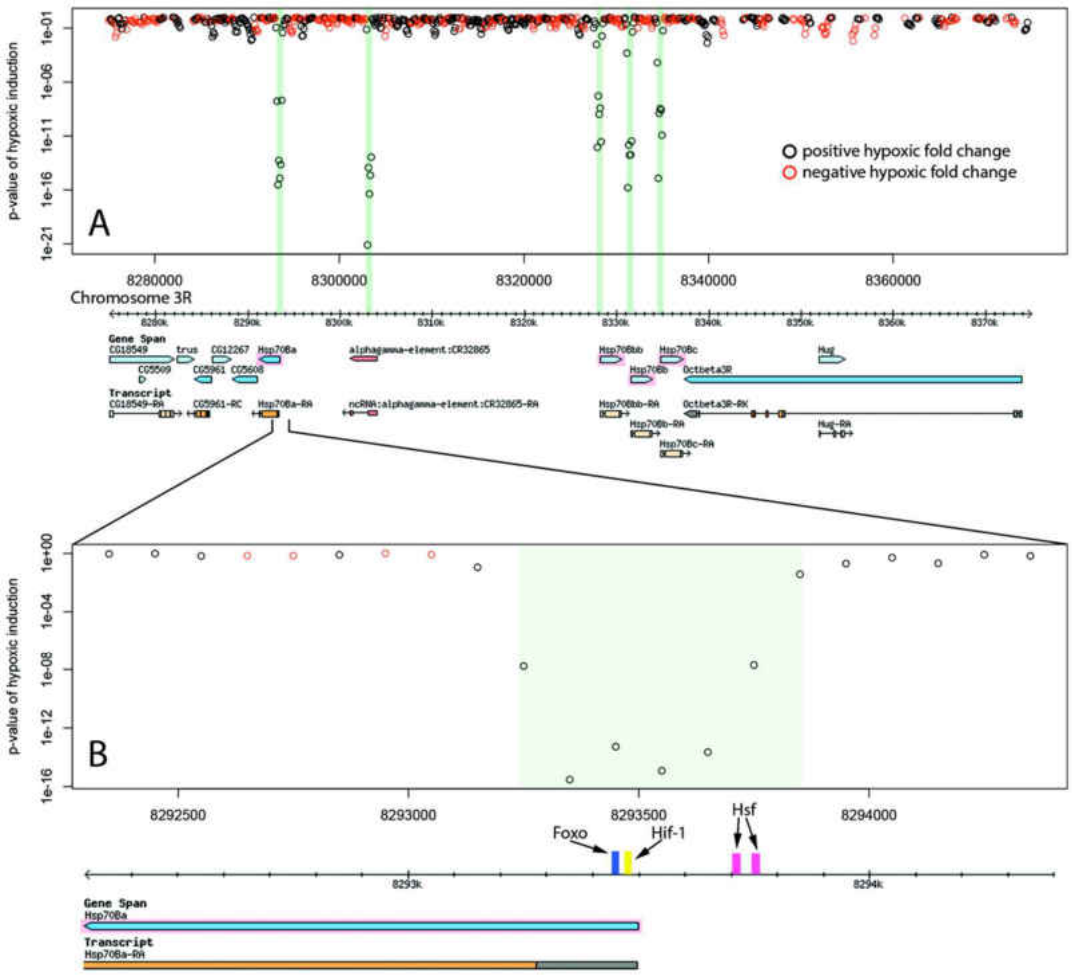

Figure 3.2. Hypoxic enhancer activity by 100bp bins at the Hsp70B locus. Each open circle plots the p-value of the difference in randomer tag counts mapping to that 100bp bin between normoxia and hypoxia. Green bars show enhancer regions discovered by our genome-wide screen. (A) The four Hsp70B homologues highlighted in pink are all up-regulated under hypoxia and contain homologous promoter proximal hypoxic enhancer regions. Additionally, a fifth homologous enhancer region lacking an ORF was discovered at the locus. (B) The close up of the Hsp70Ba enhancer region shows the position of multiple stress response transcription factor binding sites.

Six enhancers were found in introns of putative target genes (Table 3.1). These intronic enhancers may be placed proximal to alternate transcription start sites in order to confer isoform specific up-regulation as seen in the case of Sima, the Drosophila HIF-1α homologue (Figure 3.3).

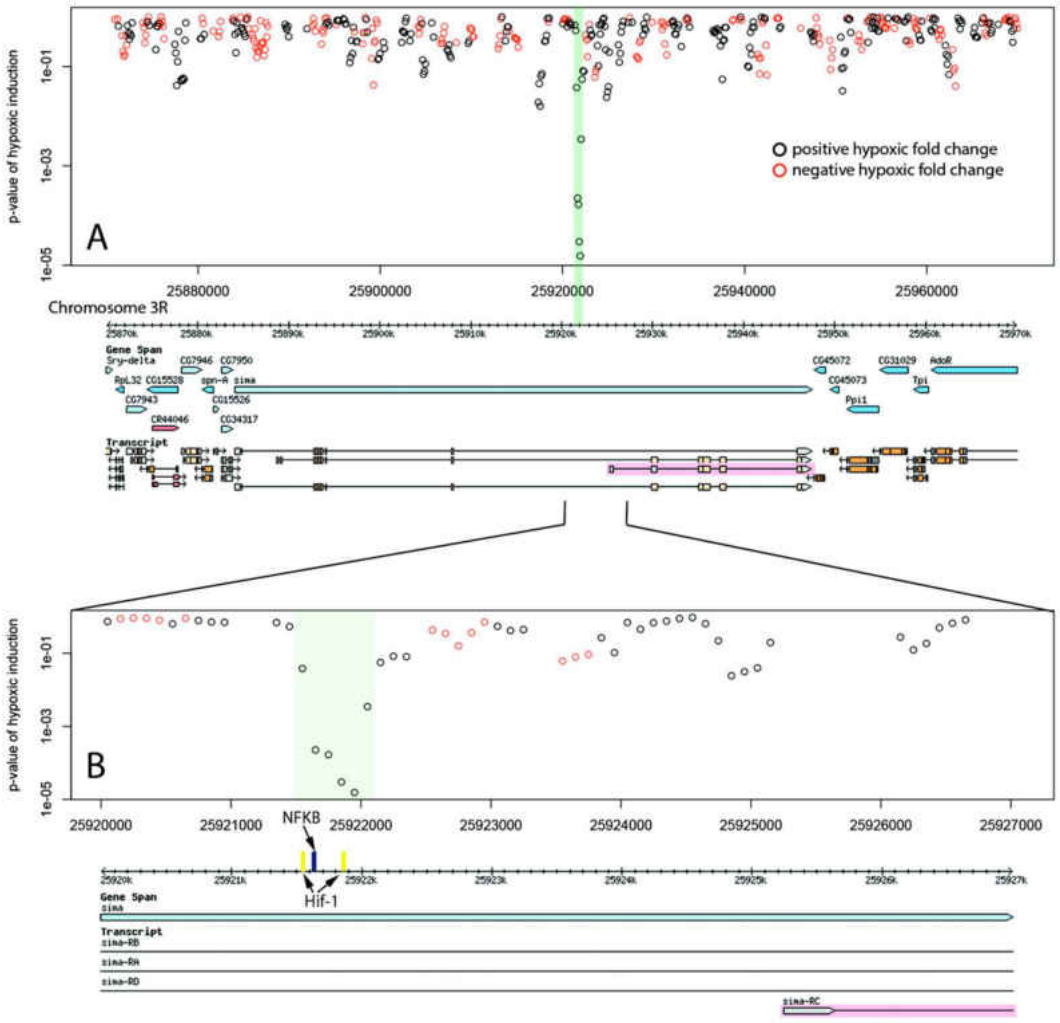

Figure 3.3. Hypoxic enhancer activity by 100bp bins at the Sima (HIF-1α) locus. Each open circle plots the p-value of the difference in randomer tag counts mapping to that 100bp bin between normoxia and hypoxia. The green bar shows the enhancer region discovered by our genome-wide screen. (A) HIF-1 is the master hypoxic regulator and is itself regulated transcriptionally under hypoxia. Our RNASeq data shows hypoxia induces up-regulation of the isoform highlighted in pink. We identify an intronic hypoxic enhancer upstream of the transcription start site of this isoform. (B) The close up of the Sima intronic enhancer region shows both HIF-1 and NF-kB binding sites.

Two enhancers were found in introns of genes neighbouring the putative target and one was found in the ORF of the putative target. The remaining five were found in intergenic space up or downstream of putative target genes, as seen for the enhancer region 13 kb downstream of the transcriptional regulator hairy (Figure 3.4).

81

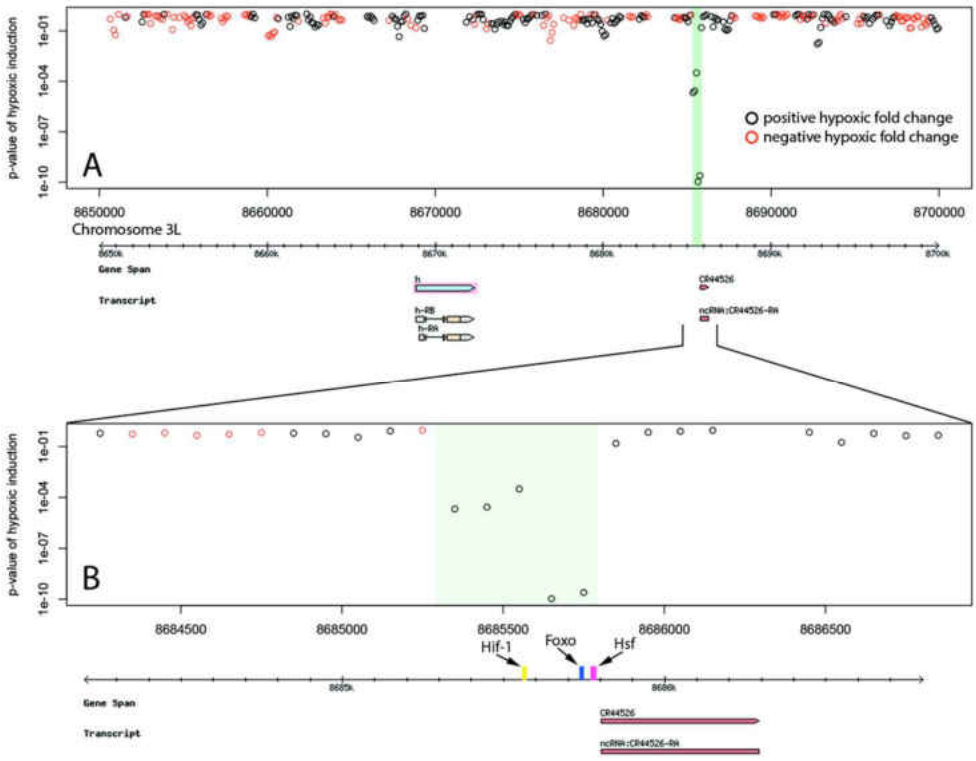

Figure 3.4. Hypoxic enhancer activity by 100bp bins at the hairy locus. Each open circle plots the p-value of the difference in randomer tag counts mapping to that 100bp bin between normoxia and hypoxia. The green bar shows the enhancer region discovered by our genome-wide screen. (A) The hairy gene produces a negative transcriptional regulator that is up-regulated during hypoxia. We identify an active hypoxic enhancer 13kb downstream of hairy. (B) The close up of the hairy downstream enhancer region shows FOXO, HIF-1 and HSF binding sites as well as coincidence with a ncRNA that is also up-regulated under hypoxia.

Interestingly, three of the five intergenic enhancers were located immediately proximal to a ncRNA. All of these ncRNAs were themselves up-regulated under hypoxia (Table 3.2).

Transcription factor binding motifs

Identified enhancer regions are enriched for binding sites of stress response transcription factors involved in hypoxia. Transcription factors HSF, HIF-1, FOXO, and NF-kB showed highly significant global enrichment across the enhancer regions (Table 3.3). Binding sites occurring in each individual enhancer are listed in Table 2.1. 26 of 31 enhancer regions contain binding motifs for at least one of these transcription factors and many contain binding sites for several. In addition to a pair of HSF binding sites, The Hsp70B promoter proximal enhancers contain binding sites for FOXO and HIF-1 (Figure 3.2). The intronic Sima enhancer (Figure 3.3) contains a pair of HIF-1 binding sites, possibly allowing autoregulation, and also contains a NF-kB binding site. The enhancer region downstream of hairy contains HSF, FOXO, and HIF-1 binding sites (Figure 3.4).

Enhancer Locus	ncRNA	Position of ncRNA relative to enhancer	Hypoxic read counts	Normoxic read counts
3R:8303000..8303500	CR32865	overlapping	66	13
3L:8685300..8685800	CR44526	3 bp upstream	31	14
3L:6256700..6257200	CR44522	201 bp upstream	6	1

Table 3.2. ncRNAs proximal to hypoxic enhancers. Three of the five enhancers not contained within protein coding transcripts coincide with ncRNAs. Each of these ncRNAs is also up-regulated under hypoxia.

Transcription Factor	P-value of Enrichment
HSF	6.22 e-12
Hif-1	6.49 e-06
Foxo	1.01 e-04
NF-kB	6.67 e-04

Table 3.3. P-value of stress transcription factor binding site enrichment in discovered enhancer sequences.

DISCUSSION

We used a novel parallelized reporter assay to conduct the first genome-wide functional enhancer screen of a cellular response to environmental stress. Our work demonstrates a new method with wide applicability and identifies DNA regulatory sequences conferring hypoxic activity. We identify 31 hypoxic enhancer regions and analyze them with respect to up-regulated hypoxic genes and stress response transcription factors.

RNA-Seq was performed on the same RNA pools used to quantify hypoxic enhancer activity in order to identify putative target genes proximal to identified enhancer regions. Differentially expressed genes identified in our RNA-Seq experiments are corroborated by previous analyses of the Drosophila hypoxic response [27,28]. The majority of enhancer regions were proximal (within 20 kb) to endogenously up-regulated genes, indicating that our enhancer assay identifies active in vivo regulatory elements. We identified enhancer regions proximal to previously described hypoxic genes including lactate dehydrogenase [6,27], the transcriptional regulator hairy [29], the reductase Wwox [30], and the cell cycle inhibitor scyl [31]. Additionally, the Hsp70B promoter proximal enhancers identified in our assay have been previously shown to be active in vivo [32,33].

The large positional overlap between up-regulated genes and enhancer regions allowed analysis of the architecture of hypoxic regulation. Interestingly, only the Hsp70B enhancers were found at the promoter of putative target genes. The majority of enhancer regions were found in introns and intergenic space. Enhancers were found in introns of putative target genes as well as introns of neighboring genes (Table 3.1). Enhancer

regions in intergenic space corresponded with known ncRNA loci and in each case the ncRNA was itself up-regulated under hypoxia (Table 3.2). These findings highlight the unbiased view of the regulatory landscape provided by genome-wide empirical assays and underscore the prevalence of activity outside of promoter regions. Some of the enhancer regions were not proximal to an identifiable up-regulated gene. These enhancers could act on more distal targets, on proximal targets with expression too low to be detected by our RNA-Seq experiment, or they may have activity in isolation but be attenuated by other elements in their native hypoxic context. Conversely, many up-regulated genes did not have a proximal enhancer identified by our screen. This could be due to a requirement of action from multiple disjunct regulatory modules at the native locus or lack of resolution in our assay. The multiple hypothesis testing correction imposed by analyzing activity across 1.2 million 100 bp bins sets a stringent p-value threshold which was not robust to noise at many loci. Genomic regions of interest can still be analyzed independently to identify enhancer activity. Future uses of the technique will benefit from further optimization of library synthesis and assay. Nonetheless, this work presents a large list of empirically identified enhancer regions robust to false discovery rate that coincide with the most highly up-regulated hypoxic genes.

The transcription factors HIF-1, HSF, NF-kB , and FOXO regulate hypoxic gene expression and have been shown to exhibit overlapping activity and reciprocal regulation [9-11, 34, 35]. The enhancer regions identified in this study are highly enriched for their binding site motifs and many display multiple sites allowing signal integration of stress response pathways. We observe an intronic enhancer in Sima which contains both HIF-1 and NF-kB binding sites, suggesting HIF-1 autoregulation and integration of NF- kB

signaling at a basal level in the hypoxic response. The enhancer region, while intronic to the full-length Sima transcript isoforms, is upstream of an alternative transcriptional start site that produces a transcript isoform that is up-regulated after hypoxia, whereas the full-length isoforms do not have altered expression after hypoxic stress. This short isoform lacks the bHLH and PAS domains of the full-length isoform, suggesting it neither binds DNA nor heterodimerizes. Interestingly, this hypoxic regulation of a short isoform resembles the hypoxic induction of a short isoform of the HIF-1 regulator fatiga (Drosophila HIF-1 Prolyl Hydroxylase) by an intronic HIF-1 enhancer [36].

Our findings reiterate the complexities of the hypoxic response while providing new details. The enhancer regions identified demonstrate regulatory activity distributed throughout non-coding genomic space and underscore the role of intronic enhancers in the hypoxic response. We observe coincidence between enhancer regions and ncRNA activity in agreement with previous evidence showing local transcription to be a general property of active enhancers [37]. We present a set of sequences capable of driving hypoxia-specific expression and demonstrate a new genome-wide technique for the identification of context-specific enhancers.

BRIDGE TO CHAPTER IV

The massively parallel enhancer assay described here has demonstrated efficacy for the analysis of transcriptional regulation in response to hypoxia. Another stress response that utilizes some of the same transcription factor pathways and binding motifs is response to bacterial infection (innate immune response). We sought to apply this high-throughput method to characterize a different biological function. In addition to

addressing a novel system in Drosophila, RNAi of transcription factors relevant to this process was included to determine their binding site specificities and interactions. The following chapter discusses this expansion on the enhancer element identification screen described here.

CHAPTER IV

ENHANCER ELEMENT IDENTIFICATION IN INNATE IMMUNE RESPONSE

This work was completed with data analysis assistance from Eric Johnson and methods development and implementation guidance from Nick Kamps-Hughes. I was the principal investigator.

INTRODUCTION

A major function of the genome is to regulate gene expression levels in response to environmental stimuli. One way this occurs is by activation of sequences known as enhancers or Cis-Regulatory Elements (CREs) which are actively bound by transcription factors to regulate the expression of nearby transcripts [1-5]. Activation of CREs varies widely by cell type even when considering a specific stress response, and there is a specific time window of activation for different stressors [2, 5-7]. In order to characterize specific responses mediated by CREs, a method is required that provides a snapshot of activity across the entire genome at a given time point.

Recent developments in high-throughput sequencing technologies have provided tools for identifying activated CREs on a genome-wide scale. Methods such as STARR-seq [5, 8] allow for quantitative readouts of enhancer activity, and are being applied to different physiological responses regulated by gene expression changes. For this investigation we use a technology developed by Nick Kamps-Hughes et al. [9] to investigate the expression of Antimicrobial Peptides (AMPs) in response to bacterial infection. This method utilizes a synthetic reporter library composed of a randomly sheared genomic fragment upstream of a minimal promoter that drives expression of a

randomer tag. Initial sequencing of a library of such constructs allows for matching of the randomer tag to its associated genomic fragment. The enhancer reporter library created for this purpose covers the Drosophila melanogaster genome at 17.39X coverage [9].

While we now possess the technology to identify CREs responding to specific stimuli on a genome-wide scale, it remains to determine what transcription factors (TFs) are functionally binding these elements. Studies have addressed the function of specific TFs at individual loci [10-14], but there is need for a comprehensive map of individual TF activity on a genome-wide scale. This would allow for comparison to determine which of a suite of candidate TFs relevant to a specific process would be most therapeutically relevant to disease outcomes due to high levels of activity.

In Drosophila melanogaster, innate immune response is one such process that is modulated by TF activity. A critical aspect of this system is its equivalency to the mammalian system [15,16], which controls expression of AMPs through the Toll and Immune Deficiency (IMD) pathways. Drosophila S2 cells are ideally suited for this investigation, as they are putative macrophages that express all pathway components for response to infection [23].

The traditional understanding of immune response pathways in Drosophila is that the IMD pathway is responsive to gram-negative bacterial infection, and activates NF-κβ TF Relish. The Toll pathway responds to gram-positive bacterial infection, and activates NF-κβ TFs Dorsal and Dorsal-related Immunity Factor (DIF) [3, 17-19]. However, there is evidence for cross-talk between the pathways, as exposure to a mixed population of bacteria can produce a greater-than-additive response [6, 20-21]. Furthermore, NF-κβ TFs DIF, Dorsal, and Relish form dimers with one another prior to binding CREs; all

possible homo- and hetero-dimers have been observed in vivo [10]. By identifying functional binding sites for each NF-κβ TF on a genome-wide scale, we hope to determine their relative contributions to innate immune response as well as characterize CREs by their binding preferences and levels of activation.

For this investigation, we employed a high-throughput enhancer screen in combination with RNAi of individual NF-κβ TFs DIF, Dorsal, and Relish. By comparing enhancer activation profiles across the genome in the presence of RNAi constructs with the same responses under no RNAi exposure, we have sought to determine the binding site preferences for each NF-κβ TF in response to different kinds of bacterial infection.

MATERIALS AND METHODS

Transfection

All assays were conducted in triplicate. Reporter library was synthesized as described in Kamps-Hughes et al. 2016 [9].

RNAi treatment groups included: nontargeting control, Dorsal, DIF, and Relish. Primers for RNAi construct amplification are as follows: Nontargeting control (from Danio rerio SlitRK exon 1) forward 5'-TACGAAGGGATTCTAGAGCAGATAC -3' reverse 5'-ACTTTCTATCTTCCTGCCCTCG-3', Dorsal forward 5'-CCGTGTATATCTCATCCAGTT-3' reverse 5'- TTGCAACAAGAGCAATATACAC-3', DIF forward 5'- TGGCACTCATTTTCTGACTTA-3' reverse 5'-GCCACAAATTGCGACCAC-3', Relish forward 5'-TCCTGTTTGTAAATTTCGAATAA-3' reverse 5'- CAGACGCCTCCGTACAAA-3'.

Drosophila melanogaster Schneider 2 cells were cultured in Schneider's Drosophila Medium (ThermoFisher) with 10% FBS. On the day of transfection, cells were split 1:2 into 3mL of growth medium in 6 well plates for 80% confluence (approximately $1x10^6$ cells per well). Transfection complex was prepared according to the product directions for FuGENE® HD Transfection Reagent (Promega) at a 3:1 ratio with 3ug reporter library DNA per well. RNAi constructs were added to transfection complex for a final concentration of 50nM.

48 hours following transfection, cells were subjected to stress in the form of peptidoglycan exposure. Gram positive peptidoglycan was from Staphylococcus aureus (InvivoGen) and gram negative peptidoglycan was from E. coli 0111:B4 (InvivoGen). Peptidoglycan was added to the wells for a final concentration of 10ug/mL. For the mixed peptidoglycans group, cells were exposed to 10ug/mL of each.

RNA isolation

RNA extraction was performed 24 hours following peptidoglycan exposure. Total RNA was extracted from cells using the Qiagen RNeasy plus mini kit according to product instructions. A minimum of 100ng of total RNA was used for RNASeq library preparation and 200ng was used for enhancer library synthesis.

Library synthesis and sequencing

For enhancer reporter library synthesis, total RNA was adjusted to a minimum of 200ng in 50uL. First strand cDNA synthesis was performed with 5 replicates per sample according to Superscript III (Thermo Fisher) product directions using random hexamer primers. The resulting cDNA was amplified using forward primer P2 (standard Illumina sequence) and a barcoded reverse primer A46 (5'-

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC GATCT-[6N barcode]-AGCCAACTTTGAATCACAAGACGCATACCAAAC-3'). Cycling parameters were as follows: 98C for 2 minutes, followed by 23 cycles of 98C for 45 seconds, 65C for 45 seconds, and 72C for 30 seconds, then cooling to 4C. Barcoded samples were then pooled and bead purified using Mag-Bind® RxnPure Plus (Omega Biotek) and size selected by running out on a 2% agarose gel and selecting for the band just under 200bp.

RNASeq libraries were prepared using the KAPA mRNA-Seq Kit according to product directions with TruSeq® barcoded adapters (Illumina).

All sequencing was performed on the Illumina HiSeq.

Data analysis

Analysis of enhancer data was conducted using the pipeline described in Kamps-Hughes et al. (2015) [9]; the only deviation from this previous analysis method was that trimming of reads to an equal number between experimental groups was conducted using the random sampling algorithm shuf rather than sampling the first reads of the files with head (as done previously). Multiple testing analyses of enhancer reporter expression levels and RNASeq expression data were performed with DESeq2.

For the purposes of identifying enhancers specific to each RNAi construct, we considered enhancer bins that showed a significant downregulation in activity in response to RNAi exposure as compared to nontargeting control with the same peptidoglycan exposure. Bins corresponding to super-enhancers were filtered from the data set by comparison to the SEA Super Enhancer Archive [22]. Significant enhancers are considered to be those with a p value of less than 0.05 and a log2 fold change of less than

-1 from matched controls (same peptidoglycan exposure under nontargeting RNAi). More stringent significance filtering greatly reduced the number of sites identified, providing a data set which contained mostly super-enhancers and did not have any significant enhancers corresponding to previously documented AMP enhancers [24]. A comparison of the number of enhancers identified per group is provided in Figure 4.1.

Scripts for DESeq2 analysis and significance filtering are provided in supplemental materials.

For motif enrichment analysis, consensus sequences were derived from Copley et al [25], Ganeson et al [26] and JASPAR regulatory element database [27]. Fasta files containing the sequences of each bin were then searched for consensus sequences using grep with the regular expressions presented in Table 4.2.

RESULTS

Overall trends in enhancer data

The total number of enhancer bins passing filters were compared between RNAi treatment groups. For the purposes of this analysis, reads were trimmed to 1,553,951 reads per replicate across all conditions. A significance threshold of $p<0.05$ with a log2 fold change<-1 was used within the DESeq2 output for the purposes of this discussion.

For all RNAi treatments, enhancers considered are those that demonstrate a significant downregulation of enhancer activity under RNAi conditions as compared to the matched nontargeting control (with the same peptidoglycan exposure). These represent enhancers that are activated by peptidoglycan exposure under normal conditions, but whose activity

is abolished in the presence of an RNAi construct. Thus, each enhancer identified is considered to be DIF, Dorsal, or Relish- responsive.

As seen in Figure 4.1, Relish-responsive enhancers are much more numerous than those responding to DIF or Dorsal in our system. This would suggest that Relish (the sole transcription factor of the IMD pathway) activates a wider set of enhancers in response to peptidoglycan exposure than the other NF-kB transcription factors. However, as will be discussed later, amplification bias of the enhancer library is a major consideration and does not allow us to make conclusions regarding relative activities of different transcription factors.

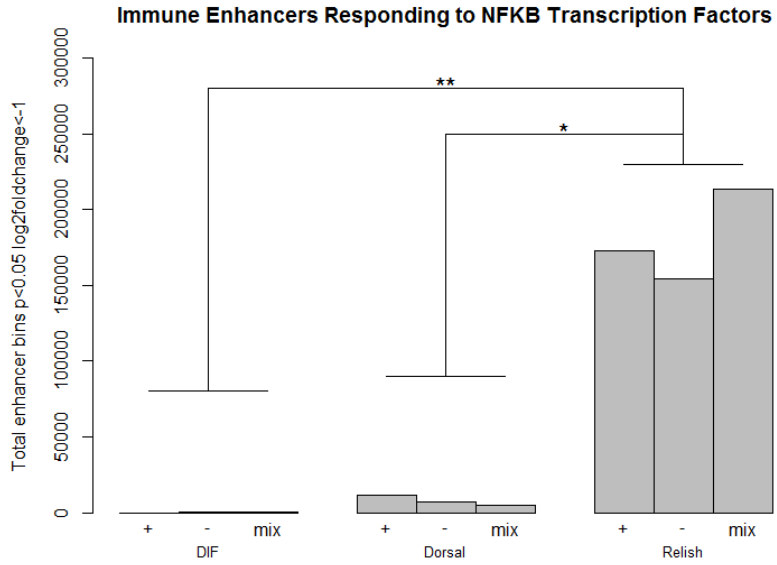

Figure 4.1. Total number of enhancers with P<0.05 and log2 fold change <-1.
Total number of enhancers passing significance filters by treatment group. Significance filters used were P value < 0.05 and log2 fold change of -1. Peptidoglycan exposure (+,-, mix) and RNAi treatment are given in the X axis labels. There is a significant difference between Relish and DIF treatment groups (P=0.006) and Relish and Dorsal treatment groups (P=0.01), but not between DIF and Dorsal treatment groups (P=0.2). Relish has significantly more enhancers than DIF or Dorsal.

Significant enhancers by peptidoglycan exposure

For the DIF RNAi group, there are approximately three-fold the significant enhancers identified for the mixed peptidoglycan exposure group than gram negative, and six-fold as compared to gram positive. This provides strong evidence that DIF is highly involved in response to mixed populations of bacteria, less active in response to gram negative bacteria, and least responsive to gram positive bacteria. This contrasts with previous studies that have found that the Toll pathway, via DIF, is responsive to gram negative bacteria [3, 6, 10, 15, 16, 18, 19]. However, it does provide evidence for cross-talk between pathways, as there is a greater-than-additive response to mixed peptidoglycans than for each individual population of either gram positive or gram negative bacteria.

Dorsal has the most significant enhancers in the gram positive peptidoglycan exposure group, which agrees with previous studies. This is followed by gram negative, and the least number of enhancers responding to Dorsal are found in the mixed peptidoglycan exposure group. This data suggests that DIF is more active than Dorsal in forming heterodimers with IMD transcription factor Relish than is Dorsal, considering their differences in response by peptidoglycan treatment.

Relish has the most significant enhancers in the mixed peptidoglycan exposure group, followed by gram negative peptidoglycan exposure. This agrees with previous studies both in increased response to mixed peptidoglycans and that the IMD pathway primarily responds to gram negative peptidoglycan [6, 10, 15, 16, 18, 19]. When considered alongside the other groups, we find that DIF and Relish are more responsive to a mixed peptidoglycan population than is Dorsal. Therefore, we conclude that DIF-

Relish heterodimers are more likely to be responding to mixed peptidoglycan exposure in our assay than are Dorsal-Relish heterodimers when considering cross-talk between pathways.

In Tanji et al. 2010 it was found that all possible NF-κβ homo- and hetero-dimers are formed in vivo, albeit at different efficiencies. The DIF-Relish heterodimer was formed at 40% efficiency, whereas that Dorsal-Relish heterodimer formed at only 7% efficiency [10]. This agrees with our finding that DIF-Relish heterodimers are more likely to respond to mixed peptidoglycans than Dorsal-Relish heterodimers.

Frequency of significant enhancer bins when omitting super enhancers

Enhancers passing significance filters were examined in relation to the SEA Super-Enhancer Archive to determine if they correspond to previously identified super enhancers. These are regions regulating gene expression that have exceptional enrichment of transcription factor binding sites, generally associated with the expression of genes that control cellular identity [22]. The frequency of enhancer bins that do not correspond to super enhancers is provided in Table 4.1, where the number of significant enhancer bins for each treatment group is provided as a percent of the total bins in the genome (each bin is 100 base pairs, therefore there are a total of 1.4 million bins in the genome). Trends between RNAi and peptidoglycan exposure groups do not differ from raw data prior to super enhancer filtering.

The percent of the genome containing DIF-responsive enhancers is very low when omitting super enhancers. None of the upstream sequences of differentially expressed genes by matched RNASeq data contain DIF-responsive enhancers passing

significance filtering. None of the upstream sequences of previously characterized AMPs contain significant DIF-responsive enhancers.

RNAi treatment	Peptidoglycan exposure	Genomic background (omitting super enhancers)	1kb upstream of all differentially expressed genes	1kb upstream of AMPs from Senger et al. 2004
DIF	+	0.0034	0	0
	-	0.0077	0	0
	mix	0.0236	0	0
Dorsal	+	0.7344	1.1628	1.2500
	-	0.4487	0.2479	0
	mix	0.2869	0	2.5000
Relish	+	10.9880	15.4264	18.7500
	-	9.7730	11.4876	6.2500
	mix	13.5396	17.4312	17.5000

Table 4.1. Frequency of significant enhancer bins by treatment group. The total number of significant enhancer bins by treatment group is presented as a percentage of total enhancer bins spanning the genome (1.4 million bins). Significant enhancers corresponding to super enhancers documented by SEA [22] are omitted. For the values corresponding to 1kb upstream of all differentially expressed genes, matched RNASeq data was used to determine genes that are upregulated in response to peptidoglycan exposure under nontargeting RNAi at a significance threshold of $P<0.05$ and log2 fold change >1. 1kb upstream of each transcription start site (TSS) was then screened for significant enhancers. 1kb upstream of the TSS of eight AMPs characterized in Senger et al. 2004 [24] was also screened for significant enhancers. Green cells indicate increased frequency as compared to genomic background; yellow cells indicate decreased frequency of significant enhancer bins.

The percent of the genome containing Dorsal-responsive enhancers is higher than that of DIF, at 0.73% for gram positive, 0.45% for gram negative, and 0.29% for mixed peptidoglycan exposure when omitting super enhancers. There is enrichment of significant enhancer bins in the 1kb upstream of differentially expressed genes in the gram-positive exposure group. In the 1kb upstream of previously documented AMPs, there is enrichment for Dorsal-responsive enhancers for gram positive as well as mixed peptidoglycan exposure. This supports our finding that Dorsal is involved in response to

gram positive peptidoglycan and stimulates expression of AMPs in response to mixed peptidoglycans.

The large number of enhancers identified as responding to Relish cover approximately 10% of the genome; this varies by peptidoglycan exposure. This is a very large proportion of the genome, but we still observed enrichment of Relish-responsive enhancers in the 1kb upstream of differentially expressed genes by matched RNASeq data sets for all peptidoglycan treatment groups. Thus, the large number of enhancers identified as being Relish-responsive are involved in differential expression. Also, there are Relish enhancers responding to all peptidoglycan exposure groups present in the 1kb upstream of AMPs, and these are enriched from the genomic background in the gram positive and mixed peptidoglycan exposure groups.

The enrichment of significant enhancers above genomic background in the 1kb upstream of previously documented AMPs in the Dorsal and Relish groups is strong evidence for the specific activity of these transcription factors for innate immune response.

Motif enrichment

Consensus motifs for each transcription factor were developed using information from Copely et al [25], Ganeson et al. [26], and JASPAR regulatory element database [27]. Some of these motifs are less stringent than others, so we cannot make comparisons between motif groups but we can draw conclusions by comparing DIF to Dorsal and Relish treatment groups.

Transcription Factor	Reference	Regular expression for consensus sequence
DIF	Copley	GG[AG][AGT][AGT][AT][ACT][ACT][CGT][CG]
Dorsal	Copley	GG[AG][AGT][ACGT][ACGT][ACT][ACT][CTG][CG]
Relish	Copley	GG[AG][AGT][AGT][ACT][CT][ACT][CTG][CGT]
DIF-Relish heterodimer	Ganeson	GGGA[AT]TC[CA]C
Dorsal	JASPAR	[CGT][GT][CTG][GT].[TGA][TA]T[TC][CTA].[CGA]
Dorsal var.2	JASPAR	[CTG][GT]G[GT][TAC][TAC]T[TC]C[CA]

Table 4.2. Consensus sequences used for motif enrichment. Represented as regular expressions.

Percent of significant enhancer bins containing consensus motifs					
		RNAi treatment			Genomic
Motif	Reference	DIF	Dorsal	Relish	background
DIF	Copley	12.96	11.73	11.66	8.80
Dorsal	Copley	23.54	22.90	22.98	18.15
Relish	Copley	18.80	16.54	16.25	12.48
DIF-Relish	Ganeson	0	0.09	0.12	0.08
Dorsal1	JASPAR	39.42	37.23	37.25	29.66
Dorsal2	JASPAR	4.74	3.76	3.92	2.82

Table 4.3. Percent of significant enhancer bins containing consensus motifs. Motifs for all transcription factors are enriched in significant enhancer bins above the genomic background, with the exception of the DIF-Relish heterodimer in the DIF-responsive enhancer group.

The consensus sequence motif for DIF homodimer binding from Copley et al [25] is enriched in all RNAi treatment groups. The group with the highest percent of bins containing this DIF binding motif is under DIF RNAi, which is as expected. Dorsal and Relish-responsive enhancers contain more DIF binding sites than the genomic background but not as many as the DIF-responsive enhancers.

The consensus binding sequence motifs for Dorsal and Relish from Copley et al [25] are enriched above the genomic background for all RNAi treatment groups. The greatest enrichment is seen in the DIF-responsive enhancer group, possibly because of the lower number of significant bins overall and the reduced amount of amplification bias in the enhancer library.

The DIF-Relish heterodimer from Ganeson et al [26] is enriched over genomic background in Dorsal and Relish-responsive enhancer bins. There are no DIF-Relish heterodimer binding motifs in the DIF-responsive enhancer group. It is possible that Relish compensates for a lack of DIF, such that a Relish homodimer might bind the DIF-Relish heterodimer consensus sequence. This is supported by the fact that the consensus binding motif sequences are overlapping between Relish homodimer and DIF-Relish heterodimer, but not for the DIF homodimer (see Table 4.2, above).

The two Dorsal binding motifs from JASPAR regulatory element database [27] are enriched above genomic background for all RNAi treatment groups. Again, we observe that there is more enrichment in the DIF RNAi group than others. Although the second Dorsal binding site consensus variant is less stringent, these frequencies are much lower than Dorsal 1 from JASPAR [27].

Selected AMP expression levels

Matched RNASeq data from the enhancer assay was used to determine the effects of loss of each transcription factor on expression levels of selected AMPs.

Cecropin A1 (CecA1) is an AMP that responds to gram negative bacterial infection via the IMD pathway [28]; however, it has also been shown to respond to gram positive bacteria [29]. Expression levels of CecA1 demonstrate a greater-than-additive effect of simultaneous exposure to mixed peptidoglycans [6]. Attacin A (AttA) is expressed in response to both gram positive and gram negative bacterial exposure, and is regulated primarily by Relish, but also by DIF to a lesser extent [10]. Drosomycin (Drs) is an antifungal peptide with a well-characterized response to both the Toll and IMD pathways, though it demonstrates severely defective induction in Toll pathway mutants

[6]. Metchnikowin (Mtk) is involved in antifungal and bacterial defense response, and responds to gram positive and gram negative peptidoglycans [30].

For the purposes of this investigation, we analyzed the effects of DIF, Dorsal, and Relish RNAi on expression levels of CecA1, AttA, Drs, and Mtk. Results are presented in Figure 4.2 as log2 fold change from nontargeting RNAi. CecA1, AttA, and Mtk show reduced expression levels under the various RNAi conditions, indicating that the transcription factors are positive regulators of expression for these genes. Drs expression levels increase under RNAi conditions at this time point, indicating that DIF, Dorsal, and Relish can act as repressors of expression for Drs.

CecA1 is regulated by DIF, Dorsal, and Relish under all conditions, but the effect of DIF RNAi in the mixed peptidoglycan exposure group is modest; it is likely that there is a compensatory effect of Dorsal or Relish in the absence of DIF in response to mixed peptidoglycans (Figure 4.2A). AttA is regulated by DIF and Dorsal under gram positive peptidoglycan exposure, a canonical Toll pathway-mediated response. Under gram negative peptidoglycans, AttA relies on Dorsal and Relish for its expression, evidence for Dorsal being involved in the IMD pathway. AttA responds to mixed peptidoglycans via the activity of Dorsal as well, and to a lesser extent Relish. It is possible that Dorsal compensates for a lack of Relish in the system by upregulating expression of AttA. Mtk is most dramatically regulated by DIF, and shows less of a response to mixed peptidoglycans. Relish is involved in expression of Mtk under gram negative and mixed peptidoglycan exposure, and Dorsal is not necessary for Mtk expression levels.

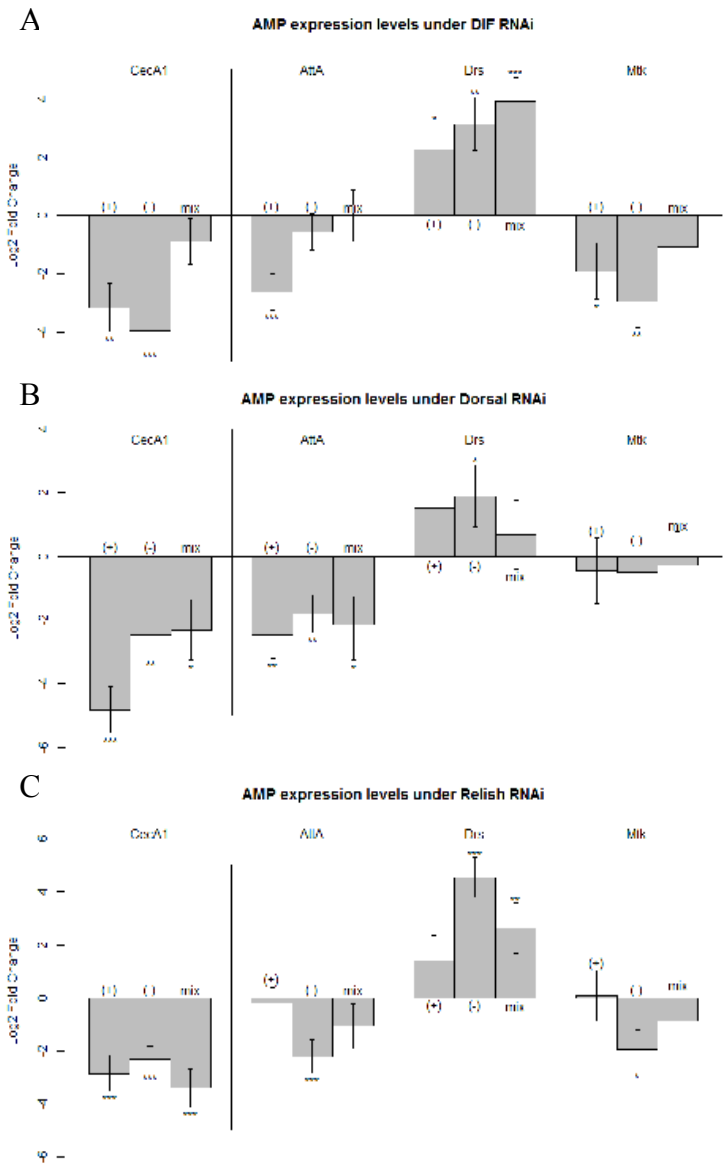

Figure 4.2. Expression levels of select AMPs under DIF, Dorsal, and Relish RNAi. Expression levels of AMPs under RNAi of DIF, Dorsal, and Relish. Values are presented as log2 fold change from nontargeting RNAi on the y axis. Peptidoglycan exposure is represented as (+)/(-)/mix on the x axis. CecA1, AttA, and Mtk show reduced expression under RNAi conditions; Drs shows increased expression.

Drs shows a dramatic increase in expression under RNAi conditions for all transcription factors tested. At this time point, it is likely that DIF, Dorsal, and Relish act as transcriptional repressors for Drs to attenuate antimicrobial response after an initial

burst of expression. RNAi of all three transcription factors causes increased expression of Drs under gram positive and negative peptidoglycan exposure; only DIF and Relish cause a significant change in expression levels under mixed peptidoglycans. Thus, it can be concluded that the DIF-Relish heterodimer is responsible for Drs expression- indicative of cross-talk between the Toll and IMD pathways.

AMP upstream enhancer activity

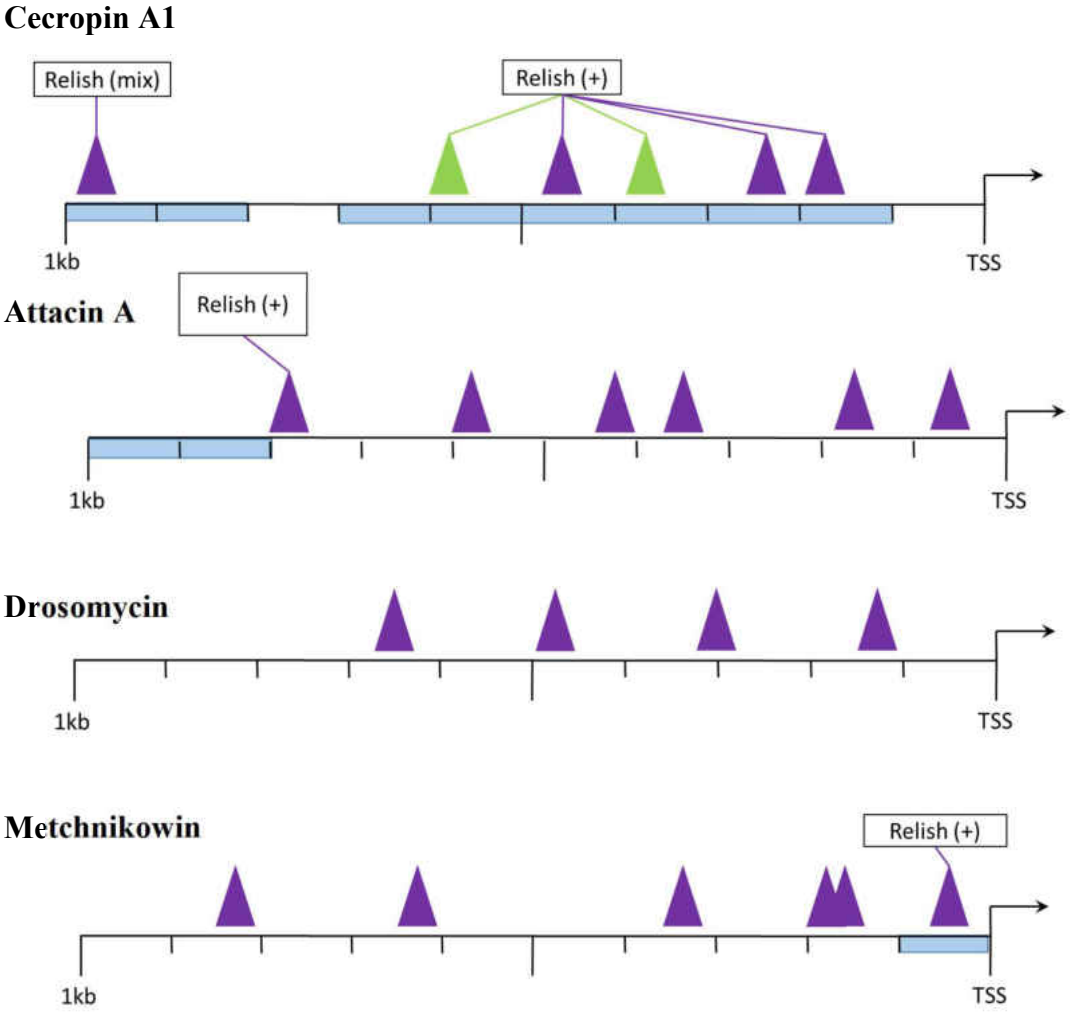

Figure 4.3. Known NF-kB sites annotated for functionally bound transcription factors in AMP promoters. Genes are from Senger et al. 2004; figure includes 1kb of sequence upstream of TSS for each. NF-kB sites are represented by purple triangles. Novel NF-kB sites identified using JASPAR insect sequence analysis and corresponding with enhancer

activity in our assay are represented by green triangles. Sites are labeled with the transcription factor identified in our screen as being necessary for site activation, as well as the peptidoglycan exposure used to stimulate enhancer activity: (+) for gram positive peptidoglycan, (-) for negative peptidoglycan, (mix) for mixed peptidoglycans. Previously documented NF-kB sites with no associated enhancer activity in our assay are not labeled with an associated transcription factor. Blue rectangles represent all significant enhancer bins by our assay.

Cecropin A1 locus

Cecropin A1 (CecA1) is one of several Cecropins found in Drosophila, with a broad antimicrobial spectrum against gram positive and gram negative bacteria [31]. It has been classified by several groups as inducible in response to gram negative bacteria and fungi [10, 12, 17, 24, 31]. Cecropins induce lysis of bacterial cells by direct disruption of their membranes, without affecting eukaryotic cells [31].

Senger et al. 2004 identified a high density of NF-kB sites immediately upstream of the TSS of CecA1 [24]. The positioning of these sites corresponds well with our enhancer activity data, and we were able to bioinformatically identify two additional NF-kB sites at the locus. We then did an in-depth investigation of the enhancer activity as it was affected by DIF, Dorsal, or Relish RNAi in tandem with analysis of matched RNASeq expression data for CecA1. The results of this analysis are presented in Figure 4.4.

There are no significant enhancer peaks responsive to DIF in our assay, but expression levels of CecA1 are significantly reduced under DIF RNAi vs. nontargeting RNAi in response to gram positive and gram negative bacteria independently. There is a non-significant reduction in the mixed peptidoglycans group under DIF RNAi. The lack of significant DIF-responsive enhancer activity does not correspond well with our matched RNASeq data, which shows reduced expression for all groups under DIF RNAi.

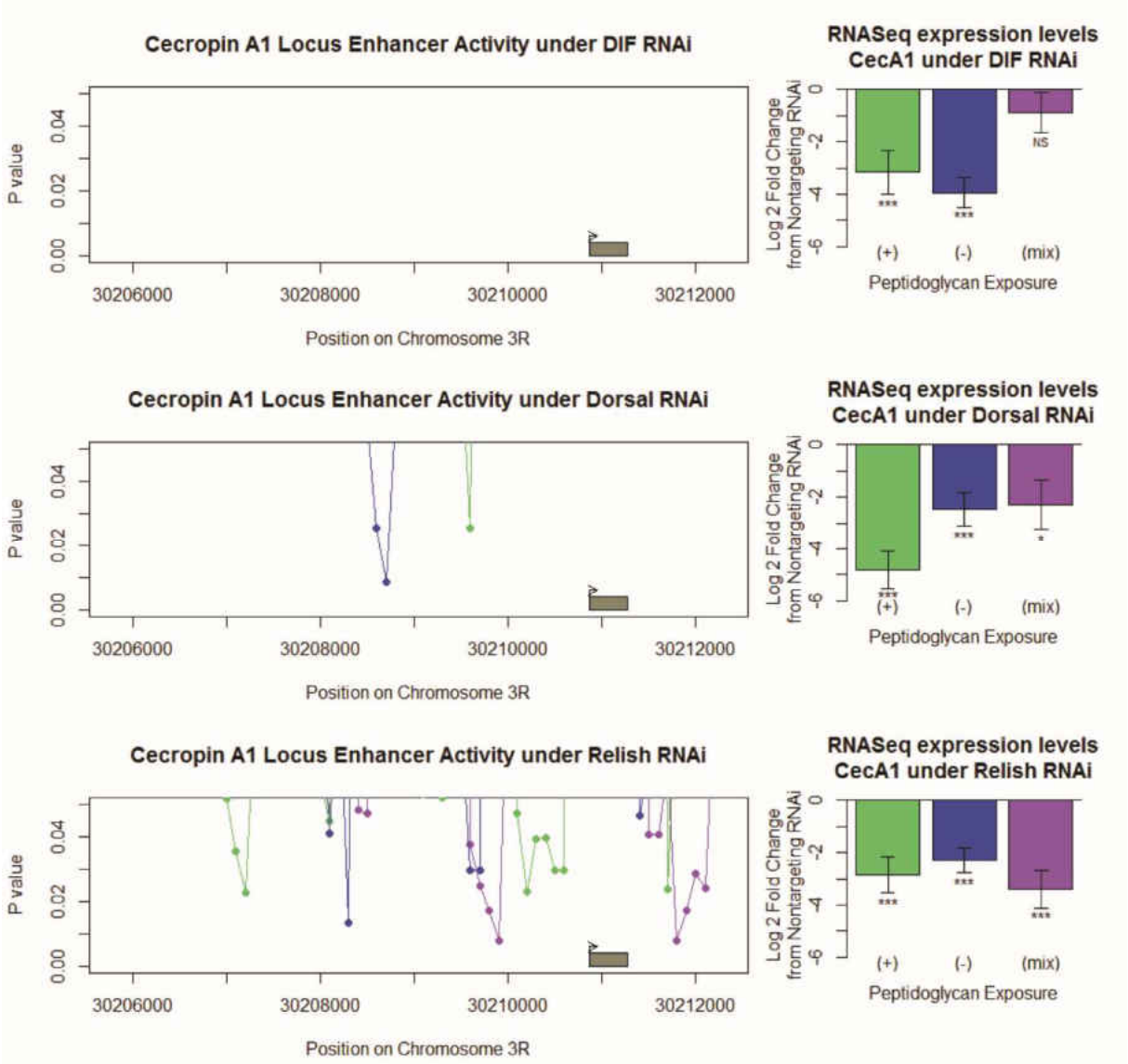

Figure 4.4. Enhancer activity and RNASeq expression levels of Cecropin A1. Enhancer Activity plots are presented in the left column for DIF, Dorsal, and Relish-responsive enhancers (top to bottom, respectively). Enhancer peaks are colored by peptidoglycan-specific response, with green representing gram positive, blue representing gram negative, and purple representing mixed peptidoglycan exposure. The x axis in these plots is the position on the chromosome, and the y axis is the significance of the enhancer activity for each 100bp bin, such that lesser P values are more significant data points and are visualized as larger peaks. The CecA1 transcript is represented in grey. Matched RNASeq data for each assay is presented in the right column. This RNASeq data is presented as log2 fold change from nontargeting RNAi, such that a significant reduction in expression is indicative of the necessity of each transcription factor for expression of CecA1 under each peptidoglycan exposure condition.

There is one significant Dorsal enhancer peak for gram positive and gram negative bacteria, but they do not overlap. Thus, different NF-kB sites are responding to Dorsal under these different pathogens. This corresponds well with our RNASeq data, which shows a significant reduction in CecA1 expression for Dorsal RNAi as compared to nontargeting control for the gram positive and gram negative peptidoglycan exposure groups. There are no significant enhancer peaks in the CecA1 upstream region responding to Dorsal in response to mixed peptidoglycan exposure, but we do see a significant reduction in expression under Dorsal RNAi that is comparable to the gram-negative group.

There is a high density of significant enhancer peaks that are responsive to Relish for all peptidoglycan exposure groups, and this corresponds well with our RNASeq data. There is a significant reduction in CecA1 expression under Relish RNAi compared to the nontargeting control for all peptidoglycan exposure groups. The upstream enhancer peaks identified colocalize with NF-kB sites identified by Senger et al 2004 [24], as well as additional sites identified in our own analysis. There are also significant Relish-responsive enhancer peaks downstream of the CecA1 transcript.

The lack of correlation between the DIF RNAi enhancer activity profile and RNASeq dataset are troubling, as is the overall increase in activity observed when examining DIF, Dorsal, and then Relish profiles. However, the data looks especially promising when examining the Relish RNAi treatment group, because of its nice correlation with both RNASeq data and NF-kB binding site locations. Noting these differences, the enhancer activity libraries used for transfection between RNAi treatment

groups were reexamined to determine if any underlying differences may be impacting our data quality.

Amplification bias of enhancer library

Only upon examining the data at high resolution and on a large scale were we able to identify differences between treatments as far as indicated by total number of significant bins (Figure 4.1), which differ significantly between Relish and the other RNAi treatment groups. This trend is especially evident when considering the CecA1 locus enhancer activity plots (Figure 4.4). There are no significant DIF-responsive enhancer bins, two significant Dorsal enhancer bins, and many significant Relish responsive enhancer bins at the CecA1 locus.

In observing these trends, we reexamined our methodology for generation and transfection of the enhancer reporter library. The assays under DIF, Dorsal, and Relish RNAi were conducted in sequence. To generate enough of the enhancer library for transfection, multiple PCR amplification rounds were conducted between assays. It is likely that PCR amplification bias resulted in the differences observed between data sets, as the DIF RNAi assay underwent the least amplification, followed by Dorsal, and finally the Relish RNAi assay, which had the most cycles of PCR amplification to generate transfection quantities of the enhancer library.

To quantify this potential PCR bias, we asked the question: how many bins are not represented in our enhancer data for each assay? If there is PCR bias affecting our data, then we would expect that certain bins would be preferentially amplified due to the inherent nature of the sequence (GC content or other factors could influence amplification bias). Thus, in the presence of PCR bias there would be more bins that are

not represented in our enhancer data because others would be overrepresented due to their preferential amplification. The results of this analysis are presented in Table 4.4, below. The transfection library used for each RNAi construct was the same for each of its peptidoglycan exposure groups (aside from random differences in transfection efficiency), so the average number of bins with zero counts is presented for each RNAi treatment.

Enhancer Bins with Zero Counts by Treatment			
RNAi Treatment	**Peptidoglycan Exposure**	**Number of Bins with Zero Counts**	**Average for RNAi treatment**
DIF	(+)	769,856	790,024
	(-)	781,514	
	mix	818,702	
Dorsal	(+)	924,828	903,430
	(-)	896,727	
	mix	888,735	
Relish	(+)	1,196,415	1,190,901
	(-)	1,171,385	
	mix	1,204,903	

Table 4.4. Enhancer bins with zero counts by treatment. For each treatment group, the number of bins with no representation in the enhancer activity dataset are presented as a measure of amplification bias in the enhancer library.

DISCUSSION

The high-throughput enhancer assay developed by Nick Kamps-Hughes et al. is an effective strategy for identifying regulatory elements on a genome-wide scale. The inclusion of RNAi constructs for transcription factors relevant to stress response was a novel facet of this project, and we have good confidence in its efficacy as we can see evidence of effective knockdown (confirmation by qPCR) as well as a difference in enhancer activity in the absence of these transcription factors. However, there was a

significant flaw in the experimental design which does not allow us to make comprehensive claims regarding our finding.

For the purposes of this investigation, we determined the presence of amplification bias in the transfection library that made our data unreliable. This investigation would need to be repeated with a newly synthesized enhancer library (thus eliminating excessive rounds of amplification to generate enough library for transfection) to provide reliable conclusions. However, the wide utility of this method, in that it can be used to analyze response to really any concieveable stressor, as well as in any model system, makes it a highly relevant technology to a multitude of investigations. In conjunction with the RNAi strategy, this method could potentially help to dissect highly complex pathways on a multitude of systems.

BRIDGE TO CHAPTER V

The enhancer assays described in this chapter and the one previous are strategies for the identification of regulatory elements in the genome that are important for stress response and rapid, real-time adaptation to changing environmental factors. As a new technology, this method provides a novel data type as its output. The next technology described allows for improvement of data quality with traditional NGS data. PELE-Seq encompasses both wet-lab and bioinformatics strategies to reduce the standard error rate inherent to short-read NGS data.

CHAPTER V

HIGH-SPECIFICITY DETECTION OF RARE ALLELES WITH PAIRED-END LOW ERROR SEQUENCING (PELE-SEQ)

This work was published in BMC Genomics on June 14, 2016 by Preston JL, Royall A, Randel MA, Sikkink KL, Phillips PC, and Johnson EA. I contributed to the development of the method by independently designing an assay to validate the presence of rare alleles identified by PELE-Seq.

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 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 5.1. 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.

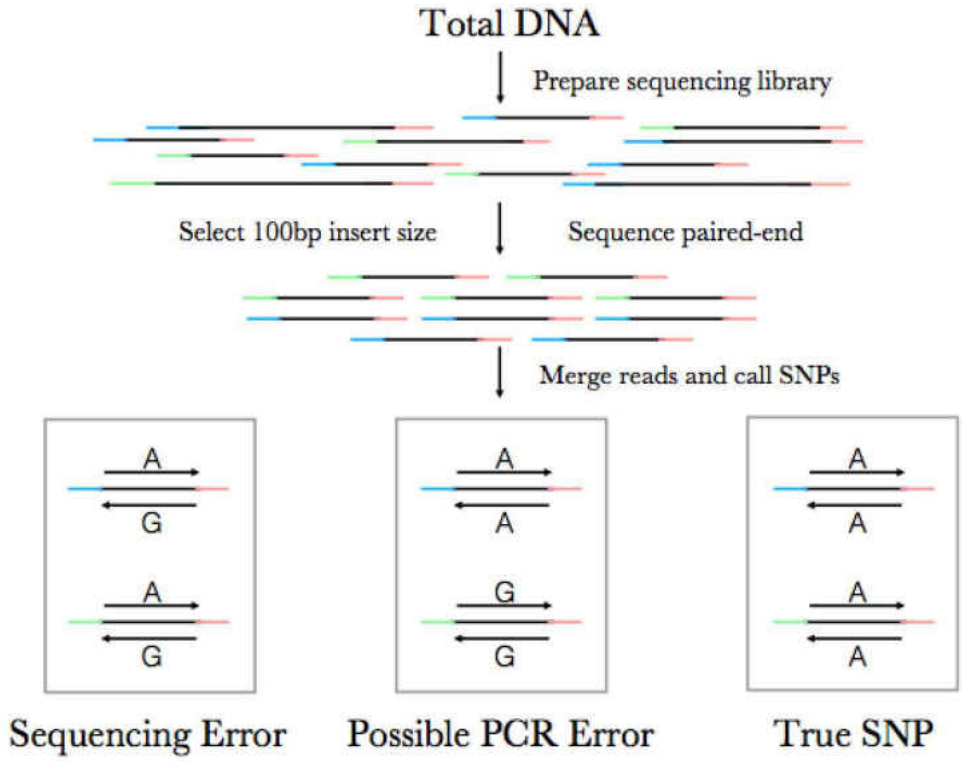

Figure 5.1. Overview of Paired-End Low Error Sequencing (PELE-Seq) library generation. DNA libraries with a 100bp insert size are paired-end sequenced using 100bp reads, generating an overlap region of approximately 100bp. The overlapping reads are merged into a consensus sequence and mismatching bases 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 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.

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 5.1). 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 5.2, 5.3). 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%.

	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 5.1. 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.

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 5.2).

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.

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 5.2. 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.

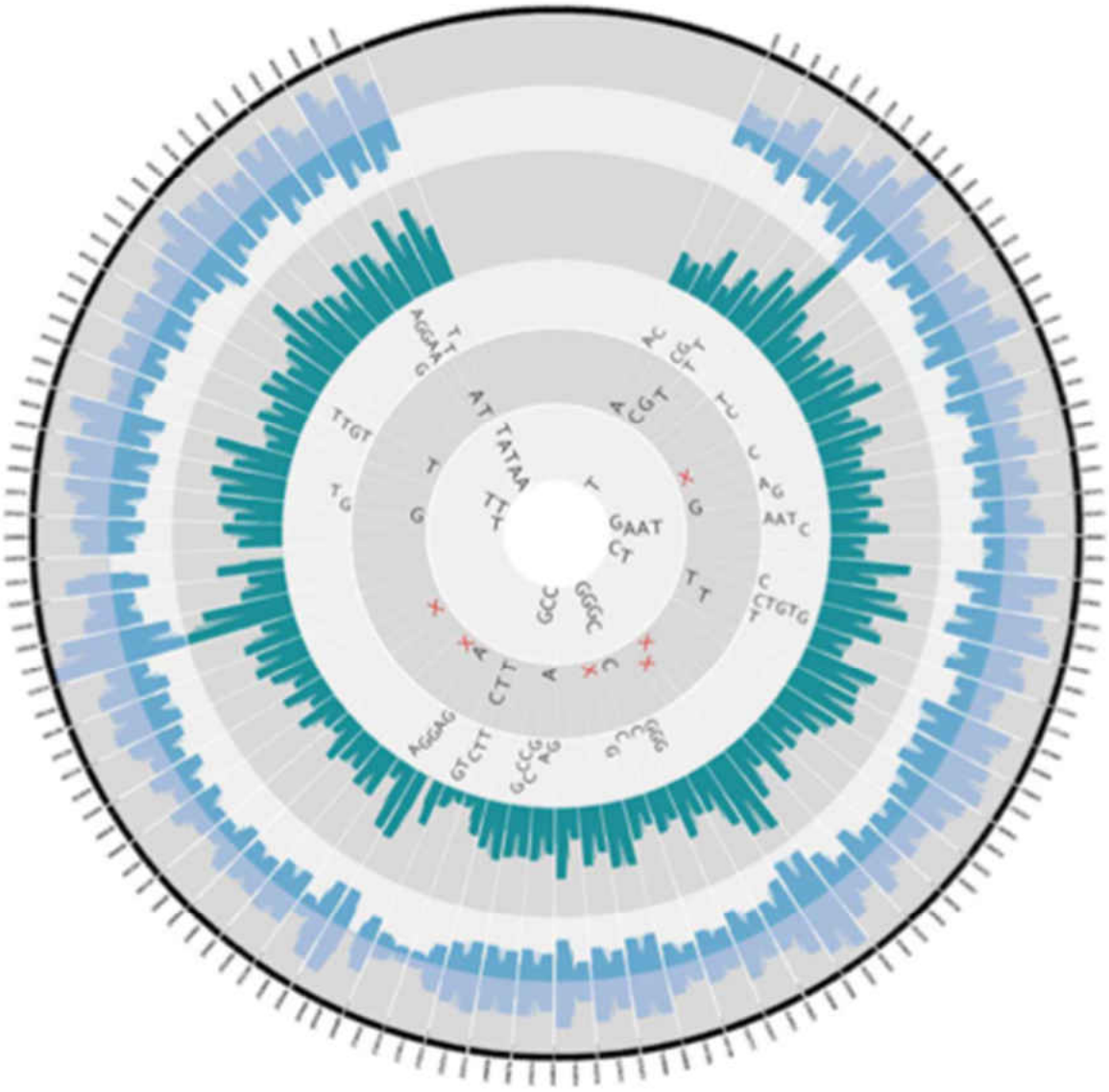

Figure 5.2. Detecting SNPs present at 0.3% frequency in E. coli control libraries with PELE-Seq and standard DNA-Seq methods at 20,000X average read depth. The read depths of the individual barcode files are plotted in blue, and the total read depth is plotted in green. The SNPs detected with PELE-Seq 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.

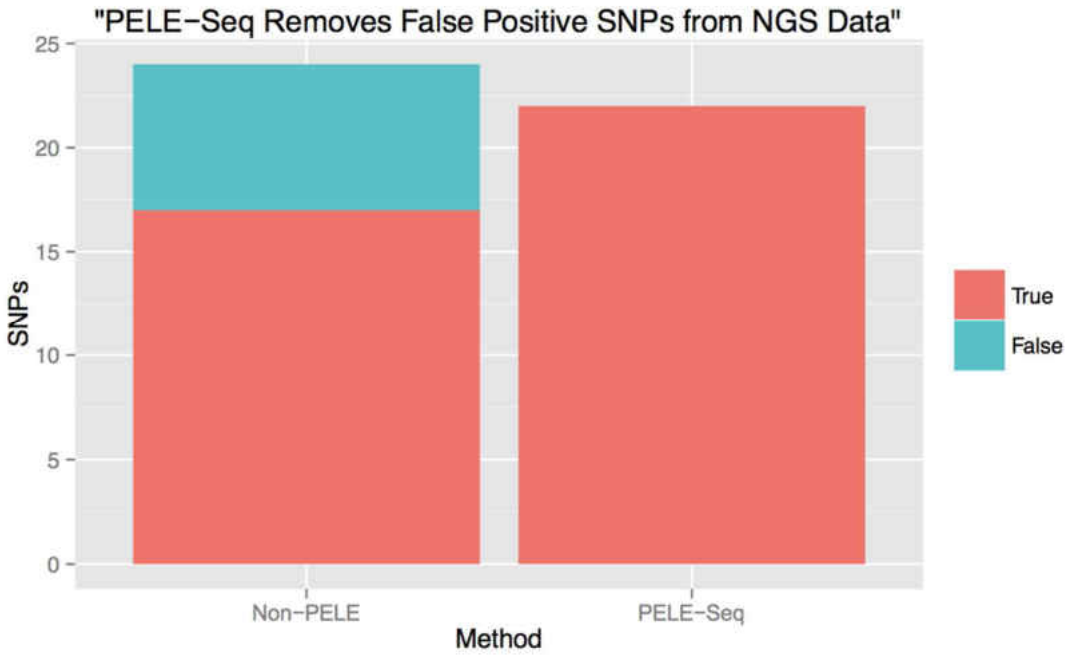

Figure 5.3. Sequencing a control E. coli DNA library containing 64 rare SNPs present at 0.3% allele frequency with PELE-Seq at 20,0000X read depth produces 100% accurate data, compared to 71% accuracy achieved with traditional sequencing methods. Traditional Non-PELE sequencing of the control libraries 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.

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 5.4). 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 5.5). 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 5.3).

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 5.3. 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 5.6). The reads mapping to this SNP in the Integrative Genome Browser (IGV) are shown in Figure 4.7.

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 5.8).

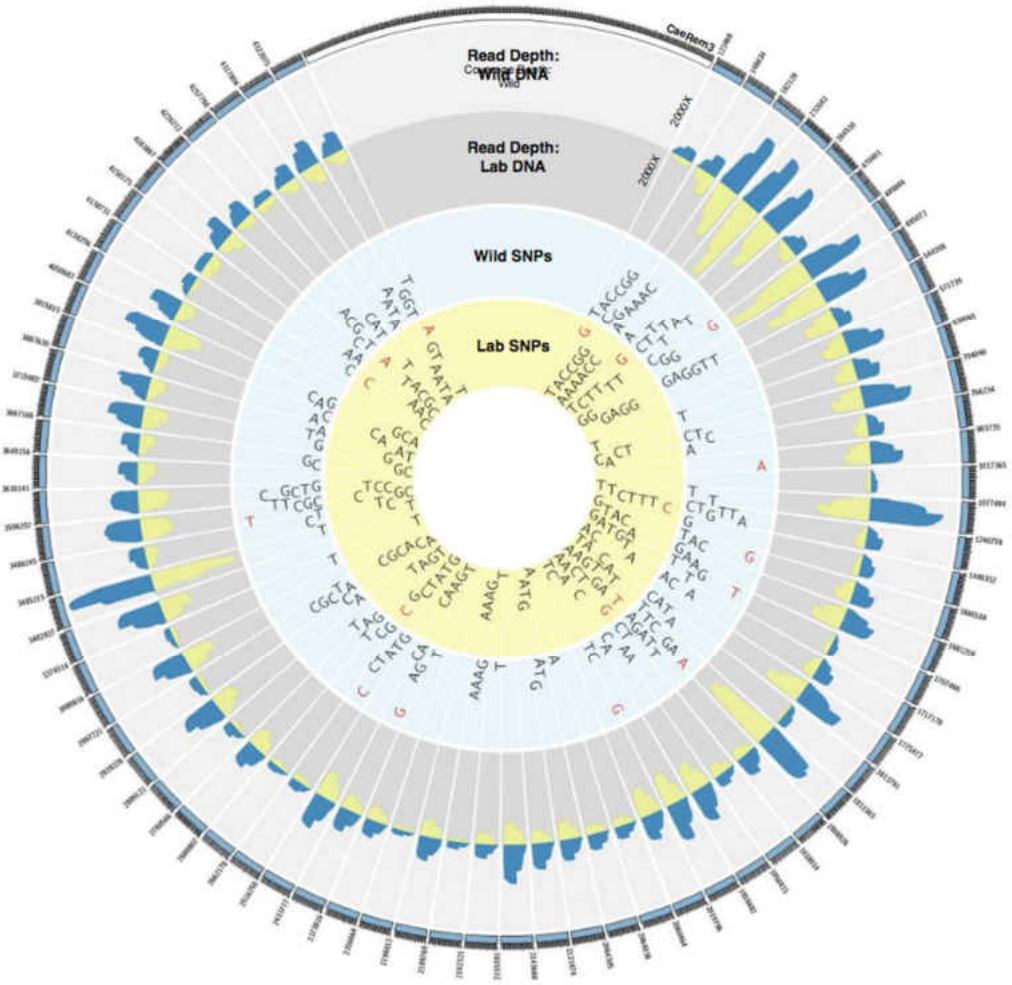

Figure 5.4. Total SNPs present in the wild and lab-adapted C. remanei populations. The inner yellow circle lists SNPs present in the lab-adapted population; the wild SNPs are listed in the blue circle. SNPs present in both the wild and lab-adapted populations are written with black letters. SNPs appearing in only the wild or lab-adapted populations are written with red letters.

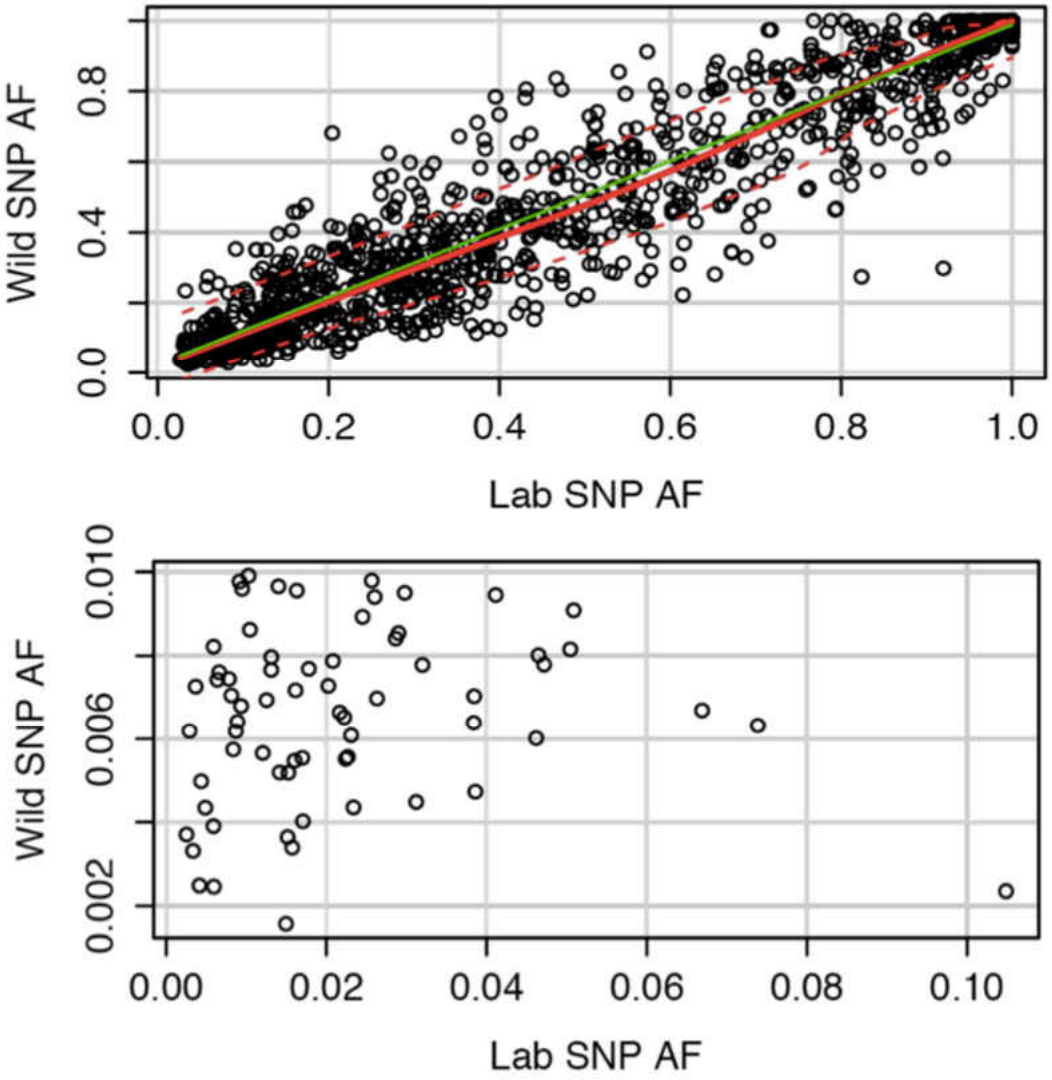

Figure 5.5. The allele frequencies of SNPs in the ancestral and lab-adapted populations of C. remanei worms. Each point represents a SNP in the genome. Top) Allele frequencies before and after lab-adaptation for all SNPs detected that are present in both populations. SNPs in the top left corner are less 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.

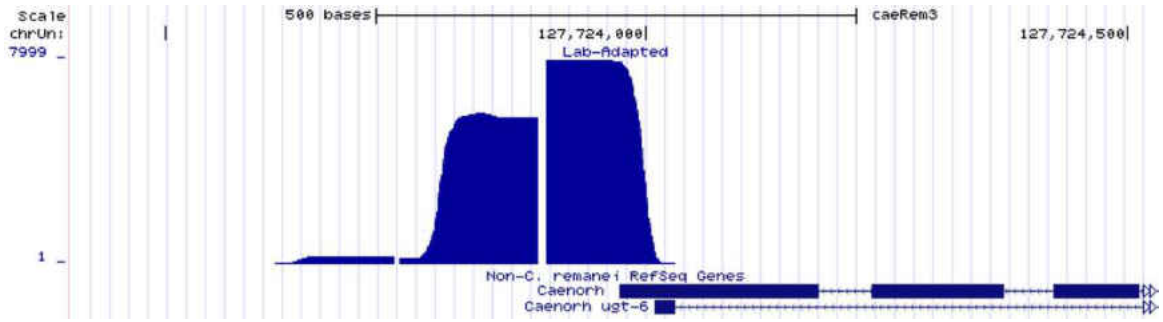

Figure 5.6. A RAD tag sequenced with PELE-Seq contains a SNP at position 127,723,967 of the caeRem3 (WUSTL) genome that maps to the predicted C. elegans gene ugt-5 that was increased in frequency by 44X after 34 generations of lab-adaptation. The UGT pathway is a major pathway responsible for the removal of drugs, toxins, and foreign substances. http://genome.ucsc.edu.

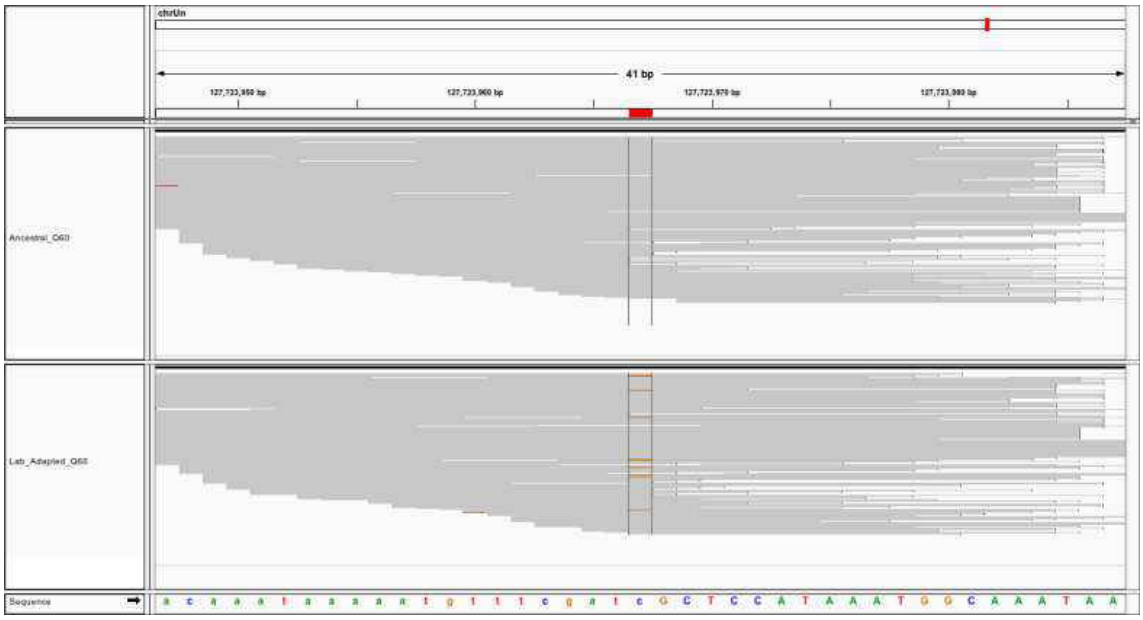

Figure 5.7. A SNP near the promoter region of ugt-5 increases in frequency 43X after lab adaptation. A G>C transversion found at below 1% frequency in the ancestral C. remanei population has a 43X increase in frequency after 34 generations of laboratory adaptation. This SNP maps to the promoter region 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.

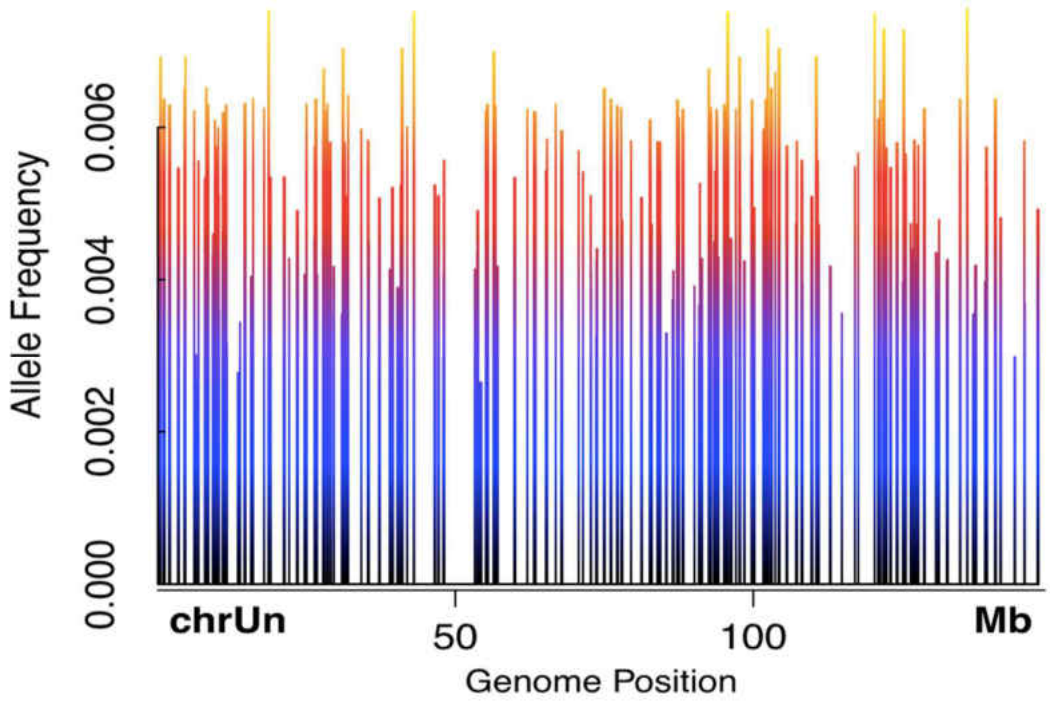

Figure 5.8. Allele frequencies and position of rare alleles detected only in the lab-adapted C. remanei population with PELE-Seq. Each vertical line represents a single SNP; the height of the line is proportional to the allele frequency. The detected SNPs had allele frequencies ranging from 0.0021 to 0.0075. The UCSC caeRem3 genome from WUSTL is composed of a single artificial chromosome named chrUn that is 146 megabases (Mb) long.

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≥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.

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' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*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. remanei 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. Thus, PELE-Seq is an effective method for improving the quality of sequencing data.

BRIDGE TO CHAPTER VI

As a novel method, PELE-Seq is effective at reducing the error rates inherent to NGS sequencing data. Other novel methods allow for the interrogation of traditional short-read sequencing in novel ways. Linked reads, described in the next chapter, provide a wealth of information inaccessible by traditional short-read sequencing including haplotype phasing, long-range information, and structural variation. In the following investigation, we apply linked read technology to identify recombination events with less investment in sequencing than ever before.

CHAPTER VI

DETECTING RECOMBINATION USING LINKED-READ TECHNOLOGY

For the purposes of this investigation, I designed and implemented an approach to apply linked-read sequencing technology in a novel way to identify recombination events. Danio rerio samples were provided by Trevor Enright. I processed these for sequencing with the assistance of Maggie Weitzman. I performed subsequent data analysis with contributions by Eric A. Johnson.

INTRODUCTION

Recombination during meiosis is a significant source of genetic diversity, allowing for the production of offspring with a combination of traits that differ from either parent. This process is required for proper assortment of haploid chromosomes, and failure can result in severe genetic abnormalities or fatality. Recombination rates differ by organism and sex; thus, characterization must be done on both sexes of each species to understand the intricacies of the process [1]. In many organisms, including humans, recombination is more likely to occur at certain loci termed ‘hot spots’ [2-5]. The identification of these hot spots is a costly challenge, requiring a large investment in sequencing given current short-read technology. This is because with a pool of DNA molecules in a traditional sequencing library, it is near impossible to determine which of the diploid chromosomes a given short read originated from. To directly resolve haplotypes, genetic variants must be identified that co-occur along a single chromosome.

Researchers have engineered creative ways to generate information about recombination frequencies. Libuda et al. have constructed fluorescent labeling strategies

for double strand breaks to visualize crossover (recombination) events. COSA-1 is a C. elegans crossover-promoting protein that forms foci at recombination sites. Only one such event occurs per chromosome in C. elegans due to profound crossover interference [7]. Using COSA-1 fluorescent labeling, they were able to determine that crossover events are associated with altered chromosome structure which inhibits additional recombination events on the same chromosome [8]. The same researchers utilized several mutants with different crossover-deficient phenotypes to establish that regulation of interhomolog interactions is a limiting factor [9]. Another strategy that does not utilize NGS technologies relies on generating gynogenetic embryos [10], which possess only genetic material from the female, as well as PCR amplification of polymorphic loci, to produce genetic and meiotic maps in Danio rerio [11-19].

We can generate a lot of information about recombination using alternative methods, but NGS technologies have not yet been developed that facilitate these investigations. In order to identify recombination events using NGS technologies, haplotypes must be characterized so that crossover events can be identified as contiguous sequences containing genetic material from each diploid chromosome. Short reads are unable to directly phase haplotypes without additional costly methods.

Such methods include: (1) haplotypes can be constructed by the sequencing of homozygous parents well as the F1 heterozygote generation [20-21]. This effectively triples the cost of analyzing a single sample. In humans, the construction of a single genetic linkage map required the genotyping of 146 families [22]. The identification of recombination events in different species now follows the same approach of sequencing or genotyping parents and the recombinant F2 progeny. The pitfalls of this approach are

that multiple generations are needed to turn the recombinant gametes into individuals, each individual has to be processed, sequenced and analyzed independently, and the precision of locating the recombination event is dependent on the marker density used to assay genetic variation.

(2) Some larger-scale projects have employed high-depth resequencing and novel bioinformatics tools to resolve haplotypes [23-24]. Typical reference genomes combine haploid sequences to form a single 'consensus' sequence that is representative of random sections of each haploid genome merged together. In cases where large rearrangements typify haplotypes or there are repeats, this can be difficult to resolve. To generate high-resolution haploid, thousands of individuals can be sequenced at high depth [25-26], as done by the 1000 Genomes Project [27] and the International HapMap Project [28]. Such methods are inherently biased by the requirement for mapping to a reference genome, as significantly divergent sequences cannot be resolved. In addition, the cost of sequencing can be astronomical in these kinds of large-scale investigations.

In this research, we asked the question: can NGS be used to cost-effectively identify recombination events? To address this question, we utilize linked-read technology for the direct detection of crossover events at a molecular level. Linked-reads and the 10x Genomics Longranger software analysis package allow for easy haplotype phasing given very low inputs (0.6-1.2ng) of sample DNA [29-31]. Linked-reads are generated by the ligation of molecule-specific barcodes to DNA within emulsion droplets, such that each DNA molecule covering a specific locus has a unique barcode identifier. A figure depicting the overall strategy for generation of linked reads is presented in Figure 6.1. Using the software package provided by 10x Genomics, we can

determine what molecule of origin each linked-read came from. This allows us to look at high resolution at haplotypes by the distribution of single nucleotide polymorphisms (SNPs) along homologous chromosomes. Haplotypes can be constructed for large regions of the genome, known as phase blocks. HMW DNA molecules containing SNPs pertaining to both haplotypes within a single phase block are indicative of recombination events.

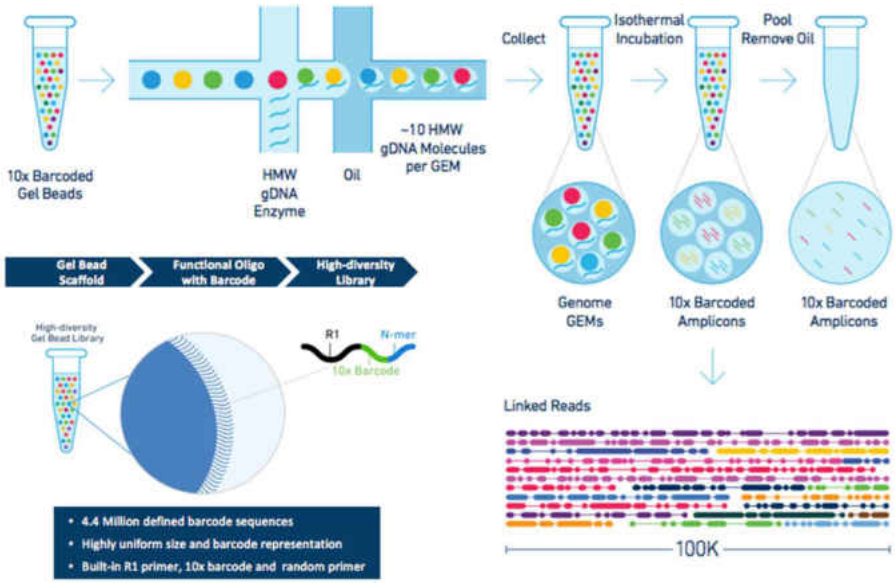

Figure 6.1. 10x Genomics linked read strategy. HMW DNA is partitioned into emulsion droplets ('GEMs') that contain a gel bead coated in barcoded oligos containing a read 1 sequencing primer ('R1'), a 16bp barcode, and a 7bp N-mer which randomly primes off of the HMW DNA molecule. This generates sequenceable fragments that are individually barcoded by their template molecule of origin. Barcodes can then be processed to generate linked read information.

The method presented here allows for identification of recombination events at an ultra-low cost of sequencing. This was accomplished by harnessing the informational power of 10x Genomics linked reads for haplotype phasing of SNPs flanking recombination sites. Even at extremely low sequencing coverage (1.23x) we were able to resolve crossover events within our data at very high resolution (within 14bp).

MATERIALS AND METHODS

Animals

Fish were maintained at 28.5C with a 14 hour light/10 hour dark cycle. The AB and Tubigen strains used in this experiment are from University of Oregon lines; more information on fish strains can be found at http://www.zfin.org.

Generation of gynogenetic embryos

We crossed two disparate Danio rerio strains (AB x Tubigen) to generate highly heterozygous F1s, then harvested sperm by gentle squeezing from anesthetized (17ppm Tricaine) males into ice cold Hank's solution. This sperm was spread thinly on a watch glass over ice, and then irradiated by a Sylvania 15W UV lamp at ~15 inches from the sample for 2 minutes while gently mixing.

We took the female F1s, squeezed their eggs, and then fertilized them with the irradiated sperm to create haploid gynogenetic embryos. Such irradiated sperm do not contribute genetic material to the embryos, therefore they contain only genetic material from the female. Haploid embryos are distinguishable by morphology after 72 hours, and will only persist for 5 days.

As these are multicellular haploid organisms, each cell within an embryo contains the same haploid genome. Therefore, there are many copies of the haploid genomes present which can be analyzed. This would not be possible if eggs themselves were harvested for this technique, as each egg would contain only one copy of the genome. Many eggs would have to be harvested, and the DNA is a very small portion of the total egg volume, complicating its isolation. For this investigation, gynogenetic embryos from a single cross were pooled for DNA extraction at 72 hours post-fertilization.

High molecular weight DNA isolation

HMW genomic DNA samples were isolated using the Qiagen HMW gDNA isolation kit according to the tissue protocol. This includes an overnight (12-16h) digestion in Proteinase K with mixing at 65C to dissolve the chorion. The resulting HMW DNA samples were then size selected for >40kB fragments on a Blue Pippin and quantified by fragment analysis to determine DNA quality. This does not give an indication of nicks, which may impact HMW DNA fragment size following denaturation.

10x Genomics Chromium platform loading strategy

In order to generate linked reads, we utilized 10x Genomics' Chromium platform. For smaller genomes, it is important to limit the number of genome copies to get resolution of individual barcoded HMW molecules without generating reads from multiple molecules at the same locus that possess the same barcode. Thus, we loaded 300 haploid genome equivalents of the Danio rerio genome with carrier DNA from other species to achieve the 0.6ng minimum loading mass.

All sequencing was performed on the Illumina HiSeq 4000.

Data analysis

The acquired sequencing data underwent all analysis prior to crossover identification using the 10x Longranger pipeline. This pipeline performs quality control filtering of reads, barcode processing (including sample indexes as well as molecule-specific barcodes), sequence alignment to a reference genome, SNP and indel calling, haplotype phasing, and large structural variant calling. Output files include industry standard bam (aligned reads sorted by barcoded molecule of origin and haplotype phasing) and vcf (variants sorted by haplotype) file types.

For identification of recombination events, a custom script was developed to first read in the phased variant genotype table in vcf format and extract the phasing information for each SNP in the table. Next, the script parsed a sorted sam file, which lists the mapping information for each read, including chromosome position and mismatches compared to the reference. If a chromosome region had multiple reads from the same 10x Chromium droplet (GEM), they were checked for phase switching, which would indicate a change from one of the diploid chromosomes to the other in a single long fragment of DNA. In order to limit false positives, only fragments that had multiple, independently sequenced loci corresponding to each phase passed the filtering.

Essentially, we filtered for linked-read molecules that have multiple consecutive SNPs pertaining to each haplotype within different short reads. Furthermore, the locus must have both haplotypes represented by linked reads spanning the crossover site, as well as multiple other linked reads supporting phasing at approximately equal representation (a heterozygous F1 parent generated the gynogenetic embryos so each haplotype will have equal representation in pooled offspring). The SNP phasing quality scores were considered in confidently calling a recombination event. These are Phred-scaled probabilities that alleles are sorted correctly in a heterozygote as compared against all other SNP calls in a phase block. A Phred quality score of 30 or greater (corresponding with 99.9% accuracy) is considered a standard cutoff in sequencing data, therefore it was utilized for high confidence calls in this investigation.

RESULTS

Summary statistics from Longranger analysis

HMW DNA from pooled gynogenetic embryos was partitioned into >400,000 emulsions for barcoding of reads generated from HMW molecules. A total of 18,246,746 reads were generated specific to this experiment, for a mean sequencing depth of 1.23. At even this low depth, more than 74% of SNPs were phased. The longest phase block (continuous stretch of sequence that is haplotype-resolved) was 103,556bp in length; the average was 8,122bp. In addition, the pipeline identified 1 large structural variant and 248 short deletions at high confidence. These summary statistics are presented in Table 6.1.

Longranger version	2.1.1
GEMs detected	402346
SNPs phased	0.743256547
Longest phase block	103556
n50 phase block	8122
Input molecule length	26152.81548
Total reads	18246746
Mean depth of coverage	1.230786715
Large SV calls	1
Short deletion calls	248

Table 6.1. Summary statistics from Longranger output. The number of GEMs detected represents the number of barcoded partitions. These summary statistics are provided as a standard output from the Longranger pipeline.

This represents extremely low-cost, low-coverage sequencing of the genome. An additional investment in sequencing would increase coverage and phasing, as well as the majority of other metrics. However, as our goal was to provide a new method that accomplishes crossover site identification at low cost, the fact that we can demonstrate

the functionality of this method with such low coverage allows for the smallest possible investment in sequencing.

Phase block length is an important consideration for the ability to identify recombination events. With longer phase blocks, it is more likely that there will be multiple SNPs phased that can provide resolution of molecules with a mix of haplotypes. The distribution of phase block lengths at the low coverage used in this experiment is presented in Figure 6.2. To increase phase block length, it is important to load extremely HMW DNA molecules onto the 10x platform. The quality of our input DNA was not as high as would be optimal, owing to the overnight proteinase K incubation required to dissolve the tough cuticle on Danio embryos. This could be due to nicks that are invisible to the Blue Pippin or degradation following size selection. In the future, the DNA isolation protocol could potentially be optimized to generate larger fragments of DNA for this purpose. Additional investment in sequencing would also increase phase block length.

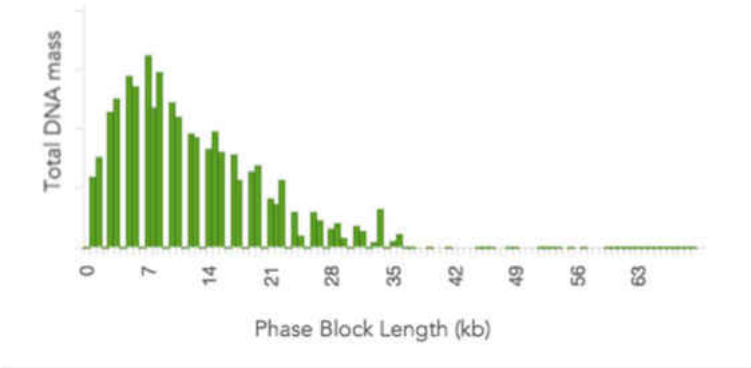

Figure 6.2. Phase block lengths generated by Longranger. Phase block lengths are presented in kilobases. Relative representation by mass within the library is presented on the y axis.

Identification of a recombination event on chromosome 6

Following the application of all filtering strategies enumerated in the methods section, a recombination event was apparent on chromosome 6. This site is spanned by linked-read molecules that represent each haplotype to allow for SNP phasing. The unphased molecule with a recombination event spans the crossover region with several kb on either side and multiple SNPs supporting each haplotype across its length. The SNP phasing quality scores at this locus are very high compared to other regions of the genome. Detailed information on the SNPs used for crossover identification is presented in Table 6.2.

This locus has relatively high linked-read coverage, though the majority of molecules are unphased. A depiction of the locus is presented in Figure 6.3. In Figure 6.3B, only the phased molecules and recombinant linked read are shown, with associated SNPs labelled by their position. The recombinant linked read is nearly 30kb in length.

Linked read molecular barcode	Position	Haplotype	Nucleotide	Phasing quality score
ACATCTTAGTTATCGC	56280525	H2	A	71
ACATCTTAGTTATCGC	56282870	H2	A	92
ACATCTTAGTTATCGC	56282884	H1	G	75
ACATCTTAGTTATCGC	56282912	H1	T	99
ACATCTTAGTTATCGC	56314446	H1	A	42

Table 6.2. Chr6: 56280525-56314446 SNPs. Each SNP originated from the same HMW DNA molecule, as evident by the shared linked read molecular barcode. Other columns list the position on chromosome 6, the haplotype of the SNP called, and the nucleotide present at that locus within the HMW molecule of origin. The phasing quality score is a Phred-like metric reported by Longranger.

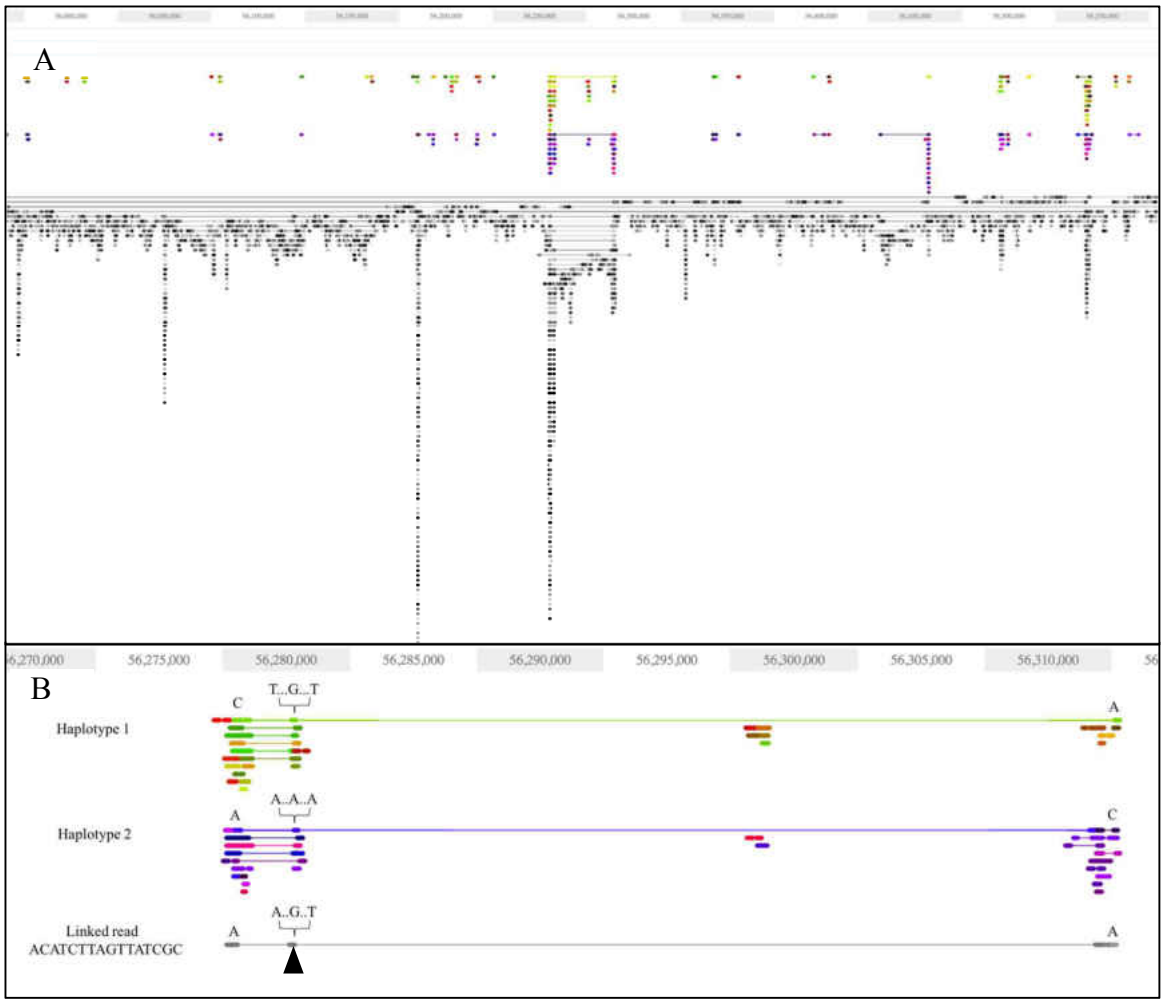

Figure 6.3. Chromosome 6 recombination locus. Two different views of the crossover locus are presented. The X axis in both plots indicates location on chromosome 6. (A) A zoomed-out view of the locus with all unphased molecules depicted in gray and phased molecules represented in color. Different shades represent different linked-read molecules. Each dot represents a short sequencing read, and lines connecting them are representative of linked reads. Haplotype 1 is shown in green/red, haplotype 2 is shown in purple/blue. (B) A zoomed-in view of the recombination site depicting only the phased molecules and the linked read containing a recombination event. SNPs are listed by location and haplotype. SNPs on the recombinant molecule are also listed; they represent a mix of haplotypes. The black triangle represents the recombination event, between SNPs 'A' and 'G'.

The two SNPs directly flanking the crossover site (an A at 56,282,870, followed by a G at 56,282,884) are only 14 nucleotides apart. Thus, we have very high resolution of this crossover site, to within 14bp. This is a resolution that is near-impossible to

achieve with traditional linkage mapping strategies. It is extremely promising that this method is able to achieve such an accurate indication of crossover location. This resolution is dependent on SNP frequencies in the organism of interest, thus it is important that there is a high level of heterozygosity which allows for the phasing of many heterozygous SNPs. Researchers should design their study such that it utilizes two highly divergent parental strains.

One thing that is evident at this locus is that there are regions with large pileups of reads (generating the 'spiky' profile in the figure 6.3A). The sequencing coverage used for this experiment is low and not uniform across the genome. This is a common phenotype in Illumina sequencing data, and should be resolved by adding more coverage. There is also a preference for certain loci within the genome relative to G/C content in Illumina data. If the observed bias were due to PCR amplification bias, then each region of high coverage would be represented by the same molecular barcode. PCR duplicates are automatically filtered by the Longranger pipeline. Thus, we still have confidence in the validity of our recombination event at this locus, owing to the high phase quality scores of SNPs used for identification.

Identification of a recombination event on chromosome 25

Another locus at which we identified a recombination event is located on chromosome 25. This site is spanned by several linked-read molecules that represent each haplotype to allow for confident SNP phasing, with higher coverage of each haplotype than the previous example. The unphased molecule with a recombination event spans the crossover region with multiple consecutive SNPs supporting each haplotype across its length.

The SNP phasing quality scores at this locus are a bit lower than those associated with the crossover event on chromosome 6, but the increased coverage of each haplotype provides high confidence that phasing was accurate at this locus. Detailed information on the SNPs used for crossover identification is presented in Table 6.3. The recombinant linked read is approximately 20kb in length.

Linked read molecular barcode	Position	Haplotype	Nucleotide	Phasing quality score
TACTGCCCATACAGAA	36804200	H2	G	41
TACTGCCCATACAGAA	36821291	H2	A	44
TACTGCCCATACAGAA	36821368	H1	C	44
TACTGCCCATACAGAA	36821768	H1	C	38
TACTGCCCATACAGAA	36821771	H1	C	38
TACTGCCCATACAGAA	36821797	H1	C	56
TACTGCCCATACAGAA	36821798	H1	T	56
TACTGCCCATACAGAA	36821830	H1	T	32
TACTGCCCATACAGAA	36821941	H1	G	25
TACTGCCCATACAGAA	36821942	H1	A	25

Table 6.3. Chromosome 25: 36804200-36821942 SNPs. Each SNP originated from the same HMW DNA molecule, as evident by the shared linked read molecular barcode. Other columns list the position on chromosome 6, the haplotype of the SNP called, and the nucleotide present at that locus within the HMW molecule of origin. The phasing quality score is a Phred-like metric reported by Longranger.

The two SNPs flanking the recombination site (A at 36,821,291 and C at 36,821,368) both have a phasing quality score of 44 and are 77bp apart. This does not provide as much resolution as the previous crossover event, but is still exceptional resolution considering our low-cost, low-coverage sequencing compared to the data resolution of alternative methods that require deep resequencing. As seen in Figure 6.4, coverage is still somewhat spiky but less dramatically so than the recombination locus on chromosome 6.

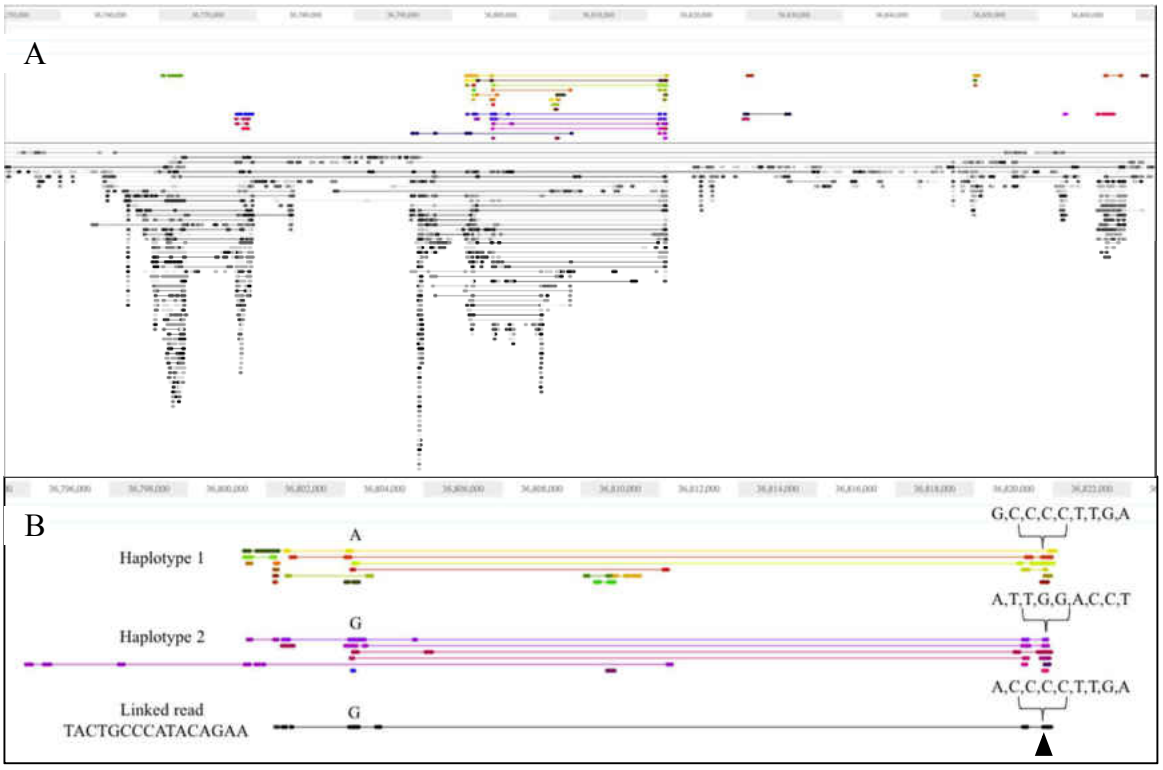

Figure 6.4. Chromosome 25 recombination locus. Two different views of the crossover locus are presented. The X axis in both plots indicates location on chromosome 6. (A) A zoomed-out view of the locus with all unphased molecules depicted in gray and phased molecules represented in color. Different shades represent different linked-read molecules. Each dot represents a short sequencing read, and lines connecting them are representative of linked reads. Haplotype 1 is shown in green/red, haplotype 2 is shown in purple/blue. (B) A zoomed-in view of the recombination site depicting only the phased molecules and the linked read containing a recombination event. SNPs are listed by location and haplotype. SNPs on the recombinant molecule are also listed; they represent a mix of haplotypes. The black triangle represents the recombination event, between SNPs 'A' and 'C' at the bottom right of the figure.

DISCUSSION

Described here is a novel method for identification of recombination events utilizing NGS without necessitating deep resequencing or population genotyping. Recombination is a vital process which allows for genetic diversity, and rates differ by organism and sex. To further our understanding of this highly variable process, many

studies are needed to characterize the intricacies of crossover regulation. Until recently, it has not been possible to phase haplotypes using NGS short-read sequencing without assuming high costs of sequencing. In addition, even these methods employ alignment to a reference genome build, which does not allow for resolution of highly divergent sequences.

The method described here could theoretically be applied to any organism that has a high level of heterozygosity, and from which haploid HMW DNA can be extracted. Gametes or gynogenesis are both good options for this. The low loading amounts (0.6-1.2ng) required for the 10x Genomics platform allow for analysis of limited samples or small organisms such as insects. In Drosophila melanogaster, which does not undergo recombination in males, F2 heterozygous females could be pooled for the purposes of this analysis. In humans, it has been shown that defects in recombination cause aneuploidy (recombination is required for proper assortment of chromosomes during meiosis) which results in infertility or inviable progeny [32]. This method could be readily applied to human sperm to determine if defective recombination is a factor in a patient's infertility. Genomes with high repeat content or complex aneuploidies may present additional challenges. Experimental design should take into consideration the quality of HMW DNA loaded, as longer DNA fragments typically generate more linked reads per molecule and therefore longer phase blocks. The number of linked reads per molecule is also impacted by sequencing depth.

In this investigation, we employed 10x Genomics linked read technology to identify recombination events by NGS at an extremely low cost of sequencing. That it is possible to identify several events at such low coverage, and achieve such fine resolution

of the loci (down to 14bp) makes this an extremely promising method. In future studies, we can work toward applying this method to determine recombination rates, identify recombination 'hot spots' and produce genetic maps. This would require more coverage than we generated, however the sequencing investment would still be significantly lower than previous methods.

CHAPTER VII

CONCLUSION

Next-Generation Sequencing (NGS) technologies are advancing at such a rapid pace that scientists are only beginning to access all the ways they may be utilized. In this investigation, we describe new methods for harnessing the potential of NGS. These include means of generating new data types, which have been radiating at a very rapid pace given the accessibility and flexibility of NGS platforms. Also described here are methods for generating data at lower cost by analyzing a consistent subset of the genome, and for addressing the error rate inherent to such large datasets. While these last methods are improving upon current technologies, they are in themselves new protocols for the application of NGS platforms.

Firstly, new data types were described to allow for the analysis of genome function in response to stress. Chapters III and IV focus on a novel method for identifying regulatory elements on a genome-wide scale that are specific to different stress responses: hypoxia and exposure to peptidoglycans (as in bacterial infection). These methods were developed to analyze the Drosophila genome, but are theoretically applicable to any organism. They utilize a reporter library for which each individual molecule contains a random fragment of the genome associated with a randomer, whose expression levels serve as a readout of enhancer activity in the genomic fragment. In chapter IV, this assay is combined with RNAi of transcription factors relevant to the pathways of interest. This genome-wide enhancer activity assay is highly accessible to

researchers developing their own investigations, as the reporter library can be generated individually for a wide variety of experiments and organisms.

Another novel data type is addressed in chapter VI, which utilizes 10x Genomics' linked read technology for the identification of recombination events. This type of assay would typically be conducted on an NGS platform by sequencing the genotypes of two homozygous parents, as well as the offspring. Thus, by utilizing this new technology for generating linked read information, we drastically reduce sequencing costs for such an investigation.

For reduction of sequencing costs, Next-RAD (as discussed in chapter II) is a method that analyzes a consistent subset of the genome between samples. This was applied to identify three separate species within the population of Anopheles darling mosquito, the primary malaria vector in Brazil. By allowing for the analysis of only a subset of the genome, many samples can be processed at the cost of sequencing one full genome per individual.

The error rate inherent to NGS has been drastically improved during the past several years but still is a major problem when dealing with such big data. PELE-Seq, described in chapter V, is a method for eliminating sequencing and PCR errors in NGS datasets. By utilizing completely overlapping paired-end reads as well as a dual barcoding strategy, we can effectively eliminate false positive low-frequency polymorphisms in our data. This method has a wide range of applications, from sequencing tumor samples to identify low-frequency drug resistance alleles, to the analysis of adaptation on a short time scale.

Finally, we presented a novel method for the identification of recombination events using linked-read sequencing. This vastly improves upon previous methods by generating data at extremely low sequencing coverage, driving down costs of such an investigation significantly. This strategy is a novel means of generating information regarding recombination on an NGS platform.

All of the technologies described here are utilizing current tools for short-read NGS in novel ways. They have only begun to be applied to the wide variety of questions that they may ultimately be able to answer. Luckily, more questions are asked of NGS technologies every day. As we continue to improve our tools, we will generate more information to assist researchers in understanding our genome and develop new therapeutic tools to benefit human health and the environment. There is much work to be done, given the rapid expansion of our toolset in this field.

REFERENCES CITED

CHAPTER I

1. Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12), 5463-7.

2. Petersen, B. S., Friedrich, B., Hoeppner, M. P., Ellinghaus, D., & Franke, A. (2017). Opportunities and challenges of whole-genome and –exome sequencing. *BMC genetics*, 18(1), 14.

3. He, K. Y., Ge, D., & He, M. M. (2017). Big data analytics for genomic medicine. *International journal of molecular sciences*, 18(2), 412.

4. Scheben A., Batley J., & Edwards D. (2016). Genotyping by sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant biotechnology journal*.

5. Buerkle C. A., & Gompert Z. (2013). Population genomics based on low coverage sequencing: how low should we go?. *Molecular Ecology*, 22(11), 3028–3035.

6. Emerson K. J., Conn J. E., Bergo E. S., Randel M. A., & Sallum M. A. (2015). Brazilian Anopheles darlingi Root (Diptera: Culicidae) clusters by major biogeographical region. *PLoS one*, 10(7), e0130773.

7. Kamps-Hughes N., Preston J. L., Randel M. A., & Johnson E. A. (2015). Genome-wide identification of hypoxia-induced enhancer regions. *PeerJ*, 3, e1527.

8. Preston J. L., Royall A. E., Randel M. A., Sikkink K. L., Phillips P. C., & Johnson E. A. (2016). High-specificity detection of rare alleles with paired-end low error sequencing (PELE-Seq). *BMC genomics*, 17(1), 464.

9. Etter P. D., Preston J. L., Bassham S., Cresko W. A., & Johnson E. A. (2011). Local De NovoAssembly of RAD Paired-End Contigs Using Short Sequencing Reads. *PLoS one*, 6(4), e18561.

10. Etter P. D., Bassham S., Hohenlohe P. A., Johnson E. A., & Cresko W. A. (2011). SNP discovery and genotyping for evolutionary genetics using RAD sequencing. *Methods in molecular biology*, 772, 157-178.

11. Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one*, *3*(10), e3376.

12. Russello, M. A., Waterhouse, M. D., Etter, P. D., & Johnson, E. A. (2015). From promise to practice: pairing non-invasive sampling with genomics in conservation. *PeerJ*, 3, e1106.

13. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one*, 7(5), e37135.

14. Franchini, P., Monné Parera, D., Kautt, A. F., & Meyer, A. (2017). quaddRAD: a new high-multiplexing and PCR duplicate removal ddRAD protocol produces novel evolutionary insights in a nonradiating cichlid lineage. *Molecular ecology*.

15. Smith, J. M., & Maynard-Smith, J. (1978) The evolution of sex. *Cambridge university press.*

16. Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C. D., ... & Aldrich, M. C. (2011). The landscape of recombination in African Americans. *Nature*, 476(7359), 170.

17. Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746), 321-4.

18. Jeffreys, A. J., Neumann, R., Panayi, M., Myers, S., & Donnelly, P. (2005). Human recombination hot spots hidden in regions of strong marker association. *Nature genetics*, 37(6), 601-6.

CHAPTER II

1. Forattini, O.P. (2002). *Culicidologia médica: identificação, biologia e epidemiologia Vol. 2.* EDUSP.

2. Castro, M. C., & Singer, B. H. (2013). Human settlement, environmental change, and frontier malaria in the Brazilian Amazon. *Ecologies and politics of health*, 118-136.

3. Hahn, M. B., Gangnon, R. E., Barcellos, C., Asner, G. P., & Patz, J. A. (2014). Influence of deforestation, logging, and fire on malaria in the Brazilian Amazon. *PLoS one*, 9(1), e85725.

4. Root, F.M. (1926). Studies on Brazilian mosquitoes. I. The Anophelines of the Nyssorhynchus group. *American journal of epidemiology*, 6(5), 684–717.

5. Galvão, A. A., Lane, J., & Corrêa, R. (1937). Notas sobre os Nyssorhynchus de São Paulo. V. Sobre os Nyssorhynchus do Novo Oriente. *Rev biol hig*, (8), 37–45.

6. Lane, J. (1939). *Catálogo dos mosquitos neotrópicos* (No. 1). Clube zoológico do Brasil.

7. Kreutzer, R. D., Kitzmiller, J. B., & Ferreira, E. (1972). Inversion polymorphism in the salivary gland chromosomes of Anopheles darlingi Root. *Mosquito news*, 32(4), 555–565.

8. Malafronte, R. S., Marrelli, M. T., & Marinotti, O. (1999). Analysis of ITS2 DNA sequences from Brazilian Anopheles darlingi (Diptera: Culicidae). *Journal of medical entomology*, 36(5), 631-634.

9. Forattini, O. P. (1987). Comportamento exófilo de Anopheles darlingi Root, em Regãio Meridional do Brasil. *Rev saude publica*, 21(4), 291–304.

10. Rosa-Freitas, M. G., Broomfield, G., Priestman, A., Milligan, P. J., Momen, H., & Molyneux, D. H. (1992). Cuticular hydrocarbons, isoenzymes and behavior of three populations of Anopheles darlingi from Brazil. *Journal of the American mosquito control association,* 8(4), 357–66.

11. Motoki, M. T., Suesdek, L., Bergo, E. S., & Sallum, M. A. (2012). Wing geometry of Anopheles darlingi Root (Diptera: Culicidae) in five major Brazilian ecoregions. *Infection, genetics and evolution*, 12(6), 1246–1252.

12. Hiwat, H., & Bretas, G. (2011). Ecology of Anopheles darlingi Root with respect to vector importance: a review. *Parasites & vectors*, 4, 177.

13. Pedro, P. M., & Sallum, M. A. M. (2009). Spatial expansion and population structure of the neotropical malaria vector, Anopheles darlingi (Diptera: Culicidae). *Biological journal of the Linnean society*, 97, 854–866.

14. Conn, J. E., Vineis, J. H., Bollback, J. P., Onyabe, D. Y., Wilkerson, R. C., & Póvoa, M. M. (2006). Population structure of the malaria vector Anopheles darlingi in a malaria-endemic region of eastern Amazonian Brazil. *The American journal of tropical medicine and hygiene*, 74(5), 798–806.

15. Manguin, S., Wilkerson, R. C., Conn, J. E., Rubio-Palis, Y., Danoff-Burg, J. A., & Roberts, D. R. (1999). Population structure of the primary malaria vector in South America, Anopheles darlingi, using isozyme, random amplified polymorphic DNA, internal transcribed spacer 2, and morphologic markers. *The American journal of tropical medicine and hygiene*, 60(3), 364–376.

16. Lounibos, L. P., & Conn, J. E. (2000). Malaria vector heterogeneity in South America. *American entomology*, 46(4), 238–249.

17. Mirabello, L., & Conn, J. E. (2006). Molecular population genetics of the malaria vector Anopheles darlingi in Central and South America. *Heredity*, 96(4), 311–321.

18. Mirabello, L., Vineis, J. H., Yanoviak, S. P., Scarpassa, V. M., Póvoa, M. M., Padilla, N., ... & Conn, J. E. (2008). Microsatellite data suggest significant population structure and differentiation within the malaria vector Anopheles darlingi in Central and South America. *BMC ecology*, 8(1), 3.

19. Alves, L. F., Vieira, S. A., Scaranello, M. A., Camargo, P. B., Santos, F. A., Joly, C. A., & Martinelli, L. A. (2010). Forest structure and live aboveground biomass variation along an elevational gradient of tropical Atlantic moist forest (Brazil). *Forest ecology and management*, 260(5), 679-691.

20. Silva, J. M. C., & Casteleti, C. H. M. (2005). Estado da biodiversidade da Mata Atlântica brasileira Mata Atlântica: Biodiversidade, Ameaças e Perspectivas. *Belo Horizonte: Fundação SOS Mata Atlântica / Conservação Internacional*.

21. Fitzpatrick, S. W., Brasileiro, C. A., Haddad, C. F., & Zamudio, K. R. (2009). Geographical variation in genetic structure of an Atlantic Coastal Forest frog reveals regional differences in habitat stability. *Molecular ecology,* 18(13), 2877–2896.

22. Morrone, J. J. (2014). Cladistic biogeography of the Neotropical region: Identifying the main events in the diversification of terrestrial biota. *Cladistics*, 30, 202–214.

23. Morrone, J. J., & Coscaron, M. C. (1996). Distributional patterns of the American Peiratinae (Heteroptera: Reduviidae). *Zool Meded (Leiden)*, 70, 1–15.

24. Angêlla, A. F., Salgueiro, P., Gil, L. H. S., Vicente, J. L., Pinto, J., & Ribolla, P. E. M. (2014). Seasonal genetic partitioning in the neotropical malaria vector, Anopheles darlingi. *Malaria journal*, 13, 203.

25. Scarpassa, V. M., & Conn, J. E. (2007). Population genetic structure of the major malaria vector Anopheles darlingi (Diptera: Culicidae) from the Brazilian Amazon, using microsatellite markers. *Memorial institute Oswaldo Cruz,* 102(3), 319–327.

26. Morrone, J.J. (2014). Biogeographical regionalisation of the Neotropical region. *Zootaxa*, 3782(1), 1–110.

27. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular ecology*, 22(11), 3124–3140.

28. Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3*, 1(3), 171–182.

29. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), R25.

30. Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959.

31. Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology*, 14(8), 2611–2620.

32. Earl, D., & vonHoldt, B. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservations genetics resources*, 4(2), 359–361.

33. Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28(24), 3326–3328.

34. Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology research*, 10(3), 564–567.

35. Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics*, 11(1), 94.

36. Jombart, T. (2008). adegenet: an R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11), 1403–1405.

37. Marinotti, O., Cerqueira, G. C., de Almeida, L. G. P., Ferro, M. I. T., Loreto, E. L. D. S., Zaha, A., ... & Pacheco, A. C. L. (2013). The genome of Anopheles darlingi, the main neotropical malaria vector. *Nucleic acids research*, 41(15), 7387-7400.

38. Meirmans, P. G. (2012). The trouble with isolation by distance. *Molecular ecology*, 21, 2839–46.

39. Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular ecology*, 22(11), 3179–3190.

40. Luca, F., Hudson, R. R., Witonsky, D. B., & Di Rienzo, A. (2011). A reduced representation approach to population genetic analyses and applications to human evolution. *Genome research*, 21(7), 1087–1098.

41. Emerson, K. J., Merz, C. R., Catchen, J. M., Hohenlohe, P. A., Cresko, W. A., Bradshaw, W. E., & Holzapfel, C. M. (2010). Resolving postglacial phylogeography using high-throughput sequencing. *PNAS*, 107(37), 16196-16200.

42. Merz, C., Catchen, J. M., Hanson-Smith, V., Emerson, K. J., Bradshaw, W. E., & Holzapfel, C. M. (2013). Replicate phylogenies and post-glacial range expansion of the pitcher-plant mosquito, Wyeomyia smithii, in North America. *PLoS one*, 8(9), e72262.

43. Gutiérrez, L. A., Gómez, G. F., González, J. J., Castro, M. I., Luckhart, S., Conn, J. E., & Correa, M. M. (2010). Microgeographic genetic variation of the malaria vector Anopheles darlingi root (Diptera: Culicidae) from Cordoba and Antioquia, Colombia. *The American journal of tropical medicine and hygiene*, 83(1), 38-47.

44. Rona, L. D., Carvalho-Pinto, C. J., Mazzoni, C. J., & Peixoto, A. A. (2010). Estimation of divergence time between two sibling species of the Anopheles (Kerteszia) cruzii complex using a multilocus approach. *BMC evolutionary biology*, 10, 91.

45. Rona, L. D., Carvalho-Pinto, C. J., & Peixoto, A. A. (2010). Molecular evidence for the occurrence of a new sibling species within the Anopheles (Kerteszia) cruzii complex in south-east Brazil. *Malaria journal*, 9(1), 33.

46. Costa, L. P. (2003). The historical bridge between the Amazon and the Atlantic Forest of Brazil: a study of molecular phylogeography with small mammals. *Journal of biogeography*, 30, 71–86.

47. Silva, D. P., Vilela, B., De Marco, P. Jr., & Nemesio, A. (2014). Using ecological niche models and niche analyses to understand speciation patterns: the case of sister neotropical orchid bees. *PLoS one*, 9(11), e113246.

48. Armstrong, K. E., Stone, G. N., Nicholls, J. A., Valderrama, E., Anderberg, A. A., Smedmark, J., ... & Richardson, J. E. (2014). Patterns of diversification amongst tropical regions compared: a case study in Sapotaceae. *Frontiers in genetics*, 5, 362.

49. Prado, C. P., Haddad, C. F., & Zamudio, K. R. (2012). Cryptic lineages and Pleistocene population expansion in a Brazilian Cerrado frog. *Molecular ecology*, 21(4), 921–941.

50. Catão, E. C., Lopes, F. A., Araújo, J. F., de Castro, A. P., Barreto, C. C., Bustamante, M., ... & Krüger, R. H. (2014). Soil acidobacterial 16S rRNA gene sequences reveal subgroup level differences between savanna-like cerrado and Atlantic forest Brazilian biomes. *International journal of microbiology*.

51. Nihei, S. S., & De Carvalho, C. J. B. (2007). Systematics and biogeography of Polietina Schnabl & Dziedzicki (Diptera, Muscidae): Neotropical area relationships and Amazonia as a composite area. *Systems entomology*, 32(3), 477–501.

52. Amorim, D., & Pires, M. (1996). Neotropical biogeography and a method for maximum biodiversity estimation. Biodiversity in Brazil: a first approach. *São Paulo: Conselho Nacional de Desenvolvimento Científico e Tecnológico*, 183–219.

53. Reidenbach, K. R., Neafsey, D. E., Costantini, C., Sagnon, N. F., Simard, F., Ragland, G. J., ... & Besansky, N. J. (2012). Patterns of genomic differentiation between ecologically differentiated M and S forms of Anopheles gambiae in West and Central Africa. *Genome biology and evolution*, 4(12), 1202-1212.

54. Cheng, C., White, B. J., Kamdem, C., Mockaitis, K., Costantini, C., Hahn, M. W., & Besansky, N. J. (2012). Ecological genomics of Anopheles gambiae along a latitudinal cline: a population-resequencing approach. *Genetics*, 190(4), 1417-1432.

55. O'Loughlin, S. M., Magesa, S., Mbogo, C., Mosha, F., Midega, J., Lomas, S., & Burt, A. (2014). Genomic analyses of three malaria vectors reveals extensive shared polymorphism but contrasting population histories. *Molecular biology and evolution*, 31(4), 889-902.

56. Cassone, B. J., Kamdem, C., Cheng, C., Tan, J. C., Hahn, M. W., Costantini, C., & Besansky, N. J. (2014). Gene expression divergence between malaria vector sibling species Anopheles gambiae and An. coluzzii from rural and urban Yaounde Cameroon. *Molecular ecology*, 23(9), 2242-2259.

CHAPTER III

1. Acevedo, J. M., Centanin, L., Dekanty, A., & Wappner, P. (2010). Oxygen sensing in Drosophila: multiple isoforms of the prolyl hydroxylase fatiga have different capacity to regulate HIFα/Sima. *PLoS one*, 5, e1527.

2. Amit, R., Garcia, H. G., Phillips, R., & Fraser, S. E. (2011). Building enhancers from the ground up: a synthetic biology approach. *Cell*, 146(1), 105-118.

3. Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10), R106.

4. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., ... & Ntini, E. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493), 455.

5. Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339, 1074–1077.

6. Arnosti, D. N., & Kulkarni, M. M. (2005). Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *Journal of cellular biochemistry*, 94, 890–898.

7. Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27, 299–308.

8. Bruick, R. K., & McKnight, S. L. (2001). A conserved family of prolyl-4-hydroxylases that modify HIF. *Science signaling*, 294, 1337–1340.

9. Bulger, M., & Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 44, 327–339.

10. Erokhin, M., Davydova, A., Parshikov, A., Studitsky, V. M., Georgiev, P., & Chetverina, D. (2013). Transcription through enhancers suppresses their activity in Drosophila. *Epigenetics chromatin*, 6(1), 31.

11. Hsu, A. L., Murphy, C. T., & Kenyon, C. (2003). Regulation of aging and age-related disease by DAF-16 and heat-shock factor. *Science*, 300(5622), 1142-1145.

12. Johnson, D., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316, 1497–1502.

13. Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T. S., & Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome research*, 23, 800–811.

14. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C., & Cohen, B. A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. *PNAS*, 109, 19498–19503.

15. Lagha, M., Bothma, J. P., & Levine, M. (2012). Mechanisms of transcriptional precision in animal development. *Trends in genetics*, 28, 409–416.

16. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9, 357–359.

17. Lavista-Llanos, S., Centanin, L., Irisarri, M., Russo, D. M., Gleadle, J. M., Bocca, S. N., ... & Wappner, P. (2002). Control of the hypoxic response in Drosophila melanogaster by the basic helix-loop-helix PAS protein similar. *Molecular and cellular biology*, 22(19), 6842-6853.

18. Li, D., Li, G., Wang, K., Liu, X., Li, W., Chen, X., & Wang, Y. (2012). Isolation and functional analysis of the promoter of the amphioxus Hsp70a gene. *Gene*, 510(1), 39-46.

19. Li, Y., Padmanabha, D., Gentile, L. B., Dumur, C. I., Beckstead, R. B., & Baker, K. D. (2013). HIF-and non-HIF-regulated hypoxic responses require the estrogen-related receptor in Drosophila melanogaster. *PLoS genetics*, 9(1), e1003230.

20. Liu, G., Roy, J., & Johnson, E. A. (2006). Identification and function of hypoxia-response genes in Drosophila melanogaster. *Physiological genomics*, 25, 134–141.

21. Ma, Q., Zhang, H., Mao, X., Zhou, C., Liu, B., Chen, X., & Xu, Y. (2014). DMINDA: an integrated web server for DNA motif identification and analyses. *Nucleic acids research*, 42, W12–W19.

22. Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., ... & Lim, J. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research*, 42(D1), D142-D147.

23. Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., ... & Kellis, M. (2012). Rapid dissection and model-based optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature biotechnology*, 30(3), 271.

24. Metzker, M.L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, 11, 31–46.

25. O'Keefe, L. V., Colella, A., Dayan, S., Chen, Q., Choo, A., Jacob, R., ... & Richards, R. I. (2011). Drosophila orthologue of WWOX, the chromosomal fragile site FRA16D tumour suppressor gene, functions in aerobic metabolism and regulates reactive oxygen species. *Human molecular genetics*, 20(3), 497-509.

26. Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., ... & Ahituv, N. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology*, 30(3), 265.

27. Perry, M. W., Boettiger, A. N., & Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *PNAS*, 108, 13570–13575.

28. Pfeiffer, B. D., Jenett, A., Hammonds, A. S., Ngo, T. T. B., Misra, S., Murphy, C., ... & Mungall, C. (2008). Tools for neuroanatomy and neurogenetics in Drosophila. *PNAS*, 105(28), 9715-9720.

29. Rius, J., Guma, M., Schachtrup, C., Akassoglou, K., Zinkernagel, A. S., Nizet, V., Johnson, R. S., Haddad, G. G., & Karin, M. (2008). NF-kappaB links innate immunity to the hypoxic response through transcriptional regulation of HIF-1alpha. *Nature*, 453, 807–811.

30. Scortegagna, M., Cataisson, C., Martin, R. J., Hicklin, D. J., Schreiber, R. D., Yuspa, S. H., & Arbeit, J. M. (2008). HIF-1α regulates epithelial inflammation by cell autonomous NFκB activation and paracrine stromal remodeling. *Blood*, 111, 3343–3354.

31. Scuderi, A., Simin, K., Kazuko, S. G., Metherall, J. E., & Letsou, A. (2006). scylla and charybde, homologues of the human apoptotic gene RTP801, are required for head involution in Drosophila. *Developmental biology*, 291, 110–122.

32. Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B. K., ... & Liu, Z. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome research*, 21(10), 1757-1767.

33. Swanson, C., Schwimmer, D. B., & Barolo, S. (2011). Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Current biology*, 21, 1186–1196.

34. Tian, S., Haney, R., & Feder, M. (2010). Phylogeny disambiguates the evolution of heat-shock cis-regulatory elements in Drosophila. *PLoS one*, 5, e1527.

35. Trapnell, C., Lior, P., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25, 1105–1111.

36. Van Uden, P., Kenneth, N. S., Webster, R., Müller, H. A., Mudie, S., & Rocha, S. (2011). Evolutionary conserved regulation of HIF-1β by NF-κB. *PLoS genetics*, 7, e1527.

37. Wang, M. C., Bohmann, D., & Jasper, H. (2005). JNK extends life span and limits growth by antagonizing cellular and organism-wide responses to insulin signaling. *Cell*, 12, 115–125.

38. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10, 57–63.

39. Wang, G.L., & Semenza, G.L. (1993). General involvement of hypoxia-inducible factor 1 in transcriptional response to hypoxia. *PNAS*, 90, 4304–4308.

40. Zhou, D., Xue, J., Lai, J. C., Schork, N. J., White, K. P., & Haddad, G. G. (2008). Mechanisms underlying hypoxia tolerance in Drosophila melanogaster: hairy as a metabolic switch. *PLoS genetics*, 4, e1527.

41. Zieler, H., & Huynh, C. Q. (2002). Intron-dependent stimulation of marker gene expression in cultured insect cells. *Insect molecular biology*, 11, 87–95.

CHAPTER IV

1. Gertz, J., Siggia, E. D., & Cohen, B. A. (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457, 215-218.

2. White, M. A. (2015). Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. *Genomics*, 106, 165-170.

3. Lemaitre, B. (2004). The road to Toll. *Nature reviews immunology*, 4, 521-527.

4. Wasserman, W. W., & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature reviews genetics*, 5, 276-287.

5. Muerdter, F., Boryn, L. M., & Arnold, C. D. (2015). STARR-seq- principles and applications. *Genomics*, 106(3), 145-150.

6. Tanji, T., Hu, X., Weber, A. N. R., & Ip, Y.T. (2007). Toll and IMD pathways synergistically activate an innate immune response in Drosophila melanogaster. *Molecular and cellular biology*, 27(12), 4578-4588.

7. Mrinal, N., & Nagarju, J. (2010). Dynamic repositioning of dorsal to two different kB motifs controls its autoregulation during immune response in Drosophila. *Journal of biological chemistry*, 285(31), 24606-24216.

8. Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, L. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339, 1074-1077.

9. Kamps-Hughes, N., Preston, J. L., Randel, M. A., & Johnson, E.A. (2015). Genome-wide identification of hypoxia-induced enhancer regions. *PeerJ*, 3, e1527.

10. Tanji, T., Yun, E.-Y., & Ip, Y. T. (2010). Heterodimers of NF-kB transcription factor DIF and Relish regulate antimicrobial peptide genes in Drosophila. *PNAS*, 107(33), 14715-14720.

11. Mrinal, N., & Nagaraju, J. (2010). Dynamic repositioning of Dorsal to two different kB motifs controls its autoregulation during immune response in Drosophila. *Journal of biological chemistry,* 285(31), 24206-24216.

12. Gross, I., Georgel, P., Kappler, C., Reichhart, J.-M., Hoffman, & J.A. (1996). Drosophila immunity: a comparative analysis of the Rel proteins dorsal and Dif in the induction of the genes encoding diptericin and cecropin. *Nucleic acids research*, 24(7), 1238-1245.

13. Malagoli, D., Accorsi, A., Sacchi, A., Basile, V., Mandrioli, M., Pinti, M., Conklin, D., & Ottaviani, E. Drosophila helical factor is an inducible protein acting as an immune-regulated cytokine in S2 cells. *Cytokine*, 58, 280-286.

14. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C., & Cohen, B. A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. *PNAS*, 109(47), 19498-19503.

15. Ganesan, S., Aggarwal, K., Paquette, N., & Silverman, N. (2011). NF-kB/Rel proteins and the humoral immune responses of Drosophila melanogaster. *Current topics microbiology immunology*, 349, 25-60.

16. Vallane, S., Wang, J. H., & Ramet, M. (2011). The Drosophila Toll signaling pathway. *Journal of immunology*, 186(2), 649-656.

17. Irving, P., Troxler, L., Heuer, T. S., Belvin, M., Kopczynski, C., Reichhart, J. M., Hoffman, J. A., & Hetru, C. (2001). A genome-wide analysis of immune responses in Drosophila. *PNAS*, 98(26), 15119-15124.

18. Busse, M. S., Arnold, C. P., Towb, P., Katrivesis, J., & Wasserman, S. A. (2007). A kB sequence code for pathway-specific innate immune responses. *The EMBO journal*, 26, 3826-3835.

19. Minakhina, S., & Steward, R. (2006). Nuclear factor-kappa B pathways in Drosophila. *Oncogene*, 25, 6749-6757.

20. Pal, S., Wu, J., & Wu, L. P. (2008). Microarray analyses reveal distinct roles for Rel proteins in the Drosophila immune response. *Developmental & comparative immunology*, 32(1), 50-60.

21. Hedengren-Olcott, M., Olcott, M. C., Mooney, D. T., Ekengren, S., Geller, B. L., & Taylor, B.J. (2004). Differential activation of the NF-kB-like factors Relish and Dif in Drosophila melanogaster by fungi and gram-positive bacteria. *Journal of biological chemistry*, 279(20), 21121-21127.

22. Wei, Y., Zhang, S., Shang, S., Zhang, B., Li, S., Wang, X., ... & Zhang, Y. (2016). SEA: a super-enhancer archive. *Nucleic acids research*, 44(D1), D172-D179.

23. Cheng, L. W., Portnoy, D. A. (2003). Drosophila S2 cells: an alternative infection model for Listeria monocytogenes. *Cellular microbiology*, 5(12), 875-885.

24. Senger, K., Armstrong, G. W., Rowell, W. J., Kwan, J. M., Markstein, M., & Levine, M. (2004). Immunity regulatory DNAs share common organization features in Drosophila. *Molecular cell*, 13, 19-32.

25. Copley, R. R., Totrov, M., Linnell, J., Field, S., Ragoussis, J., & Udalova, I. A. (2007). Functional conservation of Rel binding sites in drosophilid genomes. *Genome research*, 17(9), 1327-1335.

26. Ganesan, S., Aggarwal, K., Paquette, N., & Silverman, N. (2010). NF-κB/Rel Proteins and the Humoral Immune Responses of Drosophila melanogaster. *Current topics in microbiology and immunology*, 349, 25-60.

27. Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C. Y., Denay, G., Lee, J., ... & Zhang, A. W. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 44(D1), D110-D115.

28. Onfelt, T. T., Roos, E., & Engstrom, Y. (2001). The imd gene is required for local Cecropin expression in Drosophila barrier epithelia. *EMBO reports*, 2(3), 239-243.

29. Samakovlis, C., Kimbrell, D. A., Kylsten, P., Engstrom, A., & Hultmark, D. (1990). The immune response in Drosophila: pattern of cecropin expression and biological activity. *EMBO J*, 9, 2969-2976.

30. Meister, M., Hetru, C., & Hoffmann, J.A. (2000). The antimicrobial host defense of Drosophila. *Current topics microbiology*, 248, 17-36.

31. Steiner, H., Andreu, D., & Merrifield, R.B. (1988). Binding and action of cecropin and cecropin analogues: Antibacterial peptides from insects. *Biochimica et biophysica acta*, 939(2), 260-266.

CHAPTER V

1. Kaiser, J. (2013). The downside of diversity. *Science*, 339(6127), 1543–1545.

2. Bhatia, S., Frangioni, J., Hoffman, R., Iafrate, A. J., & Polyak, K. (2012). The challenges posed by cancer heterogeneity. *Nature biotechnology*, 30, 604–10.

3. Modi, S., Lee, H., Spina, C., & Collins, J. (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature*, 499, 219–22.

4. Hohenlohe, P., Bassham, S., Etter, P., Stiffler, N., Johnson, E. A., & Cresko, W. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS genetics*, 6(2), e1000862.

5. Nielsen, R., Paul, J. S., Albrechtsen, A., Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature reviews genetics*, 12(6), 443–451.

6. Marçais, G., Yorke, J. A., & Zimin, A. (2015). QuorUM: an error corrector for Illumina reads. *PLoS one*, 10(6), e0130821

7. Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., ... & Kota, K. (2013). Genomic variation landscape of the human gut microbiome. *Nature*, 493(7430), 45.

8. Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing - concepts and limitations. *Bioessays*, 32, 524–536.

9. Goto, H., Dickins, B., Afgan, E., Paul, I. M., Taylor, J., Makova, K. D., & Nekrutenko, A. (2011). Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. *Genome biology*, 12(6), R59.

10. Zagordi, O., Klein, R., Däumer, M., & Beerenwinkel, N. (2010). Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic acids research*, 38(21), 7400–7409.

11. Chen-Harris, H., Borucki, M., Torres, C., Slezak, T., & Allen, J. (2013). Ultra-deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs. *BMC genomics*, 14, 96.

12. Costello, M., Pugh, T. J., Fennell, T. J., Stewart, C., Lichtenstein, L., Meldrim, J. C., ... & Kim, S. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research*, gks1443.

13. Jeong, H., Barbe, V., Lee, C. H., Vallenet, D., Yu, D. S., Choi, S. H., ... & Hur, C. G. (2009). Genome sequences of Escherichia coli B strains REL606 and BL21 (DE3). *Journal of molecular biology*, 394(4), 644-652.

14. Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., ... & Horiuchi, T. (2006). Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Molecular systems biology*, 2(1).

15. Sikkink, K., Reynolds, R., Ituarte, C., Cresko, W., & Phillips, P. (2014). Rapid evolution of phenotypic plasticity and shifting thresholds of genetic assimilation in the nematode Caenorhabditis remanei. *G3: genes genomes genetics*, 4, 1103–1112.

16. Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., ... & Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic acids research*, gks918.

17. Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one*, *3*(10), e3376.

18. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., ... & Boyault, S. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463), 415.

19. Pfeifer, G. P. (2006). Mutagenesis at methylated CpG sequences. *Current topics microbiology immunology*, 301, 259–281.

20. Brodin, J., Mild, M., Hedskog, C., Sherwood, E., Leitner, L., & Andersson, B. (2013) PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. *PLoS one*, 8(7), e70388.

21. Christoforides, A., Carpten, J. D., Weiss, G. J., Demeure, M. J., Von Hoff, D. D., & Craig, D. W. (2013). Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC genomics*, 14(1), 302.

22. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H.E. (2012). Double digest RADseq: an inexpensive method for De novo SNP discovery and genotyping in model and non-model species. *PLoS one*, 7(5), e37135.

23. Pan, L., Shah, A. N., Phelps, I. G., Doherty, D., Johnson, E. A., & Moens, C. B. (2015). Rapid identification and recovery of ENU-induced mutations with next-generation sequencing and Paired-End Low-Error analysis. *BMC genomics*, 16(1), 83.

24. Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature reviews genetics*, 13, 135–145.

25. De La Vega, F. M., Bustamante, C. D., & Leal, S. M. (2011). Genome-wide association mapping and rare alleles: from population genomics to personalized medicine. *Pacific symposium biocomputing*, 74–75.

26. King, C. D., Rios, G. R., Green, M. D., & Tephly, T. R. (2000). UDP-Glucuronosyltransferases. *Current drug metabolism*, 19, 143–161.

27. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., ... & Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9), 1639-1645.

28. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, 29(1), 24.

29. Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief bioinformatics*, 14, 178–192.

30. Phanstiel, D. H. (2015). Sushi: tools for visualizing genomics data. *R package version 1.8.0*.

31. Wickham, H. (2009). ggplot2: elegant graphics for data analysis. *New York: Springer-Verlag.*

CHAPTER VI

1. Smith, J. M., & Maynard-Smith, J. (1978). The evolution of sex. *Cambridge university press*.

2. Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C. D., ... & Aldrich, M. C. (2011). The landscape of recombination in African Americans. *Nature*, 476(7359), 170.

3. Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746), 321-324.

4. Jeffreys, A. J., Neumann, R., Panayi, M., Myers, S., & Donnelly, P. (2009). Human recombination hot spots hidden in regions of strong marker association. *Nature genetics*, 37(6), 601-606.

5. Reiter, L., Murakami, T., Koeuth, T., Pentao, L., Muzny, D., Gibbs, R., & Lupski, J. (1998). A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. *Nature genetics*, 19(3), 288-297.

6. Kitzman, J. O. (2016). Haplotypes drop by drop. *Nature biotechnology*, 34(3), 296-298.

7. Martinez-Perez, E., & Colaiacovo, M. P. (2009). Distribution of meiotic recombination events: talking to your neighbors. *Current opinions genetics*, 19, 105–112.

8. Libuda, D. E., Uzawa, S., Meyer, B. J., & Villeneuve, A. M. (2013). Meiotic chromosome structures constrain and respond to designation of crossover sites. *Nature*, 502(7473), 703-706.

9. Rosu, S., Libuda, D. E., & Villeneuve, A. M. (2011). Robust crossover assurance and regulated interhomolog access maintain meiotic crossover number. *Science*, 334(6060), 1286-1289.

10. Corley-Smith, G. E., Lim, C. J., & Brandhorst, B. P. (1996). Production of androgenetic zebrafish (Danio rerio). *Genetics*, 142(4), 1265-1276.

11. Postlethwait, J. H., Johnson, S. L., Midson, C. N., Talbot, W. S., Gates, M., Ballinger, E. W., Africa, D., Andrews, R., Carl, T., Eisen, J. S., & Horne, S. A genetic linkage map for the zebrafish. *Science*, 699-703.

12. Johnson, S. L., Gates, M. A., Johnson, M., Talbot, W. S., Horne, S., Baik, K., Rude, S., Wong, J. R., & Postlethwait, J. H. (1996). Centromere-linkage analysis and consolidation of the zebrafish genetic map. *Genetics*, 142(4), 1277-1288.

13. Woods, I. G., Kelly, P. D., Chu, F., Ngo-Hazelett, P., Yan, Y. L., Huang, H., Postlethwait, J. H., & Talbot, W.S. (2000). A comparative map of the zebrafish genome. *Genome research*, 10(12), 1903-1914.

14. Singer, A., Perlman, H., Yan, Y., Walker, C., Corley-Smith, G., Brandhorst, B., & Postlethwait, J. (2002). Sex-specific recombination rates in zebrafish (Danio rerio). *Genetics*, 160(2), 649-657.

15. Johnson, S. L., Gates, M. A., Johnson, M., Talbot, W. S., Horne, S., Baik, K., Rude, S., Wong, J. R., & Postlethwait, J. H. (1996). Centromere-linkage analysis and consolidation of the zebrafish genetic map. *Genetics*, 142(4), 1277-1288.

16. Gates, M. A., Kim, L., Egan, E. S., Cardozo, T., Sirotkin, H. I., Dougan, S. T., ... & Talbot, W. S. (1999). A genetic linkage map for zebrafish: comparative analysis and localization of genes and expressed sequences. *Genome research*, 9(4), 334-347.

17. Barbazuk, W. B., Korf, I., Kadavi, C., Heyen, J., Tate, S., Wun, E., ... & Johnson, S. L. (2000). The syntenic relationship of the zebrafish and human genomes. *Genome research*, 10(9), 1351-1358.

18. Kelly, P. D., Chu, F., Woods, I. G., Ngo-Hazelett, P., Cardozo, T., Huang, H., ... & Johnson, S. L. (2000). Genetic linkage mapping of zebrafish genes and ESTs. *Genome research*, 10(4), 558-567.

19. Woods, I. G., Kelly, P. D., Chu, F., Ngo-Hazelett, P., Yan, Y. L., Huang, H., Postlethwait, J. H., & Talbot, W. S. (2000). A comparative map of the zebrafish genome. *Genome research*, 10(12), 1903-1914.

20. Steen, R. G., Kwitek-Black, A. E., Glenn, C., Gullings-Handley, J., Van Etten, W., Atkinson, O. S., ... & Granados, M. (1999). A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. *Genome Research*, *9*(6), AP1-AP8.

21. Dietrich, W. F., Miller, J., Steen, R., & Merchant, M. A. (1996). A comprehensive genetic map of the mouse genome. *Nature*, 380(6570), 149.

22. Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., ... & Shlien, A. (2002). A high-resolution recombination map of the human genome. *Nature genetics*, 31(3), 241.

23. Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18, 1851–1858.

24. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9), 1297-1303.

25. Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., ... & Sigurdsson, G. T. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature genetics*, 47(5), 435.

26. Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., ... &

Sato, Y. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature communications*, *6*.

27. 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68.

28. International HapMap Constortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature,* 449, 851-862.

29. Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., ... & Mudivarti, P. A. (2016). Haplotyping germline and cancer genomes using high-throughput linked-read sequencing. *Nature biotechnology*, *34*(3), 303.

30. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome research*, *27*(5), 757-767.

31. Hui, W. W., Jiang, P., Tong, Y. K., Lee, W. S., Cheng, Y. K., New, M. I., ... & Chiu, R. W. (2017). Universal haplotype-based noninvasive prenatal testing for single gene diseases. *Clinical chemistry*, *63*(2), 513-524.

32. Gonsalves, J., Sun, F., Schlegel, P. N., Turek, P. J., Hopps, C. V., Greene, C., ... & Pera, R. A. R. (2004). Defective recombination in infertile men. *Human molecular genetics*, *13*(22), 2875-2883.