MASSIVELY PARALLEL SEQUENCING-BASED ANALYSES OF GENOME AND PROTEIN FUNCTION

by

NICHOLAS KAMPS-HUGHES

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 2015

DISSERTATION APPROVAL PAGE

Student: Nicholas Kamps-Hughes

Title: Massively Parallel Sequencing-Based Analyses of Genome and Protein Function

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

Bruce Bowerman	Chairperson
Eric Johnson	Advisor
Tory Herman	Core Member
Kryn Stankunas	Core Member
John Conery	Core Member
J. Andrew Berglund	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 2015

DISSERTATION ABSTRACT

Nicholas Kamps-Hughes

Doctor of Philosophy

Department of Biology

June 2015

Title: Massively Parallel Sequencing-Based Analyses of Genome and Protein Function

The advent of high-throughput DNA and RNA sequencing has made possible the assay of millions of nucleic acid molecules in parallel. This allows functional genomic elements to be identified from background in single-tube experiments. This dissertation discusses the development of two such functional screens as well as work implementing a third that was previously developed in my thesis laboratory.

Restriction-Associated DNA sequencing (RAD-Seq) is a complexity reduction sequencing method that allows the same subset of genomic sequence to be read across multiple samples. Differences in sample collection and data analysis allow manifold applications of RAD-Seq. Here we use RAD-Seq to identify mutant genes responsible for altered phenotypes in *Caenorhabditis elegans* and to identify hyper-invasive alleles in trout population admixtures.

Apart from acquiring genomic sequence data, massively-parallel sequencing can be used for counting applications that quantify activity across a large number of test molecules. This dissertation describes the development of a technique for simultaneously quantifying the activity of a restriction enzyme across all possible DNA substrates by

linking digest of a sequenced genome to Illumina-sequencing in an unbiased fashion. Finally, a powerful approach to analyze transcriptional activation is described. This method quantifies output from millions of potential DNA transcriptional enhancers via RNA amplicon sequencing of covalently-linked randomer tags and is used in conjunction with RNA-Seq to provide a mechanistic view of hypoxic gene regulation in *Drosophila*.

This dissertation includes previously published, co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Nicholas Kamps-Hughes

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
University of California, Berkeley

DEGREES AWARDED:

Doctor of Philosophy, Biology, 2015, University of Oregon
Bachelor of Science, Genetics and Plant Biology, 2006, University of California

AREAS OF SPECIAL INTEREST:

Genomics
Molecular Biology

PUBLICATIONS:

Kamps-Hughes N, Quimby A, Zhu Z, Johnson EA (2013) Massively parallel characterization of restriction endonucleases. *Nucleic Acids Res* 41: e119-e119.

Hohenlohe PA, Day MD, Amish SJ, Miller MR, **Kamps-Hughes N**, Boyer MC, Muhlfeld CC, Allendorf FW, Johnson EA, Luikart G (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Mol Ecol* 22: 3002-3013.

Senn H, Ogden R, Cezard T, Gharbi K, Iqbal Z, Johnson EA, **Kamps-Hughes N**, Rosell F, McEwing R (2013) Reference-free SNP discovery for the Eurasian beaver from restriction site–associated DNA paired-end data. *Mol Ecol* 22: 3141-3150.

Rego EH, Shao L, Macklin JJ, Winoto L, Johansson GA, **Kamps-Hughes N**, Davidson MW, Gustafsson MG (2012) Nonlinear structured-illumination microscopy with a photoswitchable protein reveals cellular structures at 50-nm resolution. *PNAS* 109: E135-E143.

O'Rourke SM, Yochem J, Connolly AA, Price MH, Carter L, Lowry JB, Turnbull DW, **Kamps-Hughes N**, Stiffler N, Miller MR, Johnson EA, Bowerman B (2011) Rapid mapping and identification of mutations in Caenorhabditis elegans by restriction site-associated DNA mapping and genomic interval pull-down sequencing. *Genetics* 189: 767-778.

ACKNOWLEDGMENTS

I would like to give special thanks to Eric Johnson for providing valuable guidance and insight and Bruce Bowerman for excellent collaboration. I am grateful for the stellar colleagues Jessica Preston, Melissa Randel and Ariel Royale as well as the thoughtful input of committee members Kryn Stankunas, Tory Herman and John Conery. A big thanks to Paul Etter, Doug Turnbull, Jason Carriere and Nick Stiffler for invaluable assistance with DNA library preparation and sequencing. Big-ups to Mike D and Oggie for the S2 connection. Kudos to the University of Oregon biology department teaching staff for providing excellent graduate teaching fellowships. I would also like to acknowledge the family and friends who have helped me achieve my goals as a scientist.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF FIGURES

LIST OF TABLES

Table Page

Chapter III

Chapter IV

Chapter V

CHAPTER I

INTRODUCTION

High-throughput sequencing and the genomics era

The advance of the genomics discipline and the advent of high-throughput nucleic acid assays are intertwined in a feedback loop between theory and methodology. Beginning with microarray technology in the 1990s[1], genome-wide methods vastly increased the scope of biological experimentation. Microarrays provided quantitative transcriptome-wide information which allowed the *de novo* identification of differentially expressed genes[2,3] and helped to elucidate regulatory networks[4,5]. Researchers were able to adapt the power of microarrays toward other goals as well. In the case of RAD[6], microarrays were used to quickly identify DNA markers between populations.

High-throughput sequencing took these analyses to new heights of power and inference. By definition, microarrays can only bind and report on nucleic acid probes that were expressly printed upon them. Library preparations for sequencing instruments such as Illumina involve unbiased adapter ligations allowing all molecules present in the reaction to ultimately be sequenced. Indeed these instruments were first used for the acquisition of new genomic sequence[7,8,9] and were paramount to the assembly of numerous genomes[10,11,12]. While this remains a principle application of massively parallel sequencing platforms, they have been adapted for a range of analyses. One such application is RAD-Seq, created here at University of Oregon[13]. This technique entails the sequencing of DNA adjacent to restriction sites allowing the same widely-distributed genomic locations to be probed across samples. Differences in these sequences can then

be used to compare individuals and populations. It can be used to generate markers and map mutations in recombinant animals as described in Chapter II. This work in *C. elegans* was completed in collaboration with O'Rourke SM, Yochem J, Connolly AA, Price MH, Carter L, Lowry JB, Turnbull DW, Stiffler N, Miller MR, Johnson EA, and Bowerman B. The polymorphic tags generated can also be used to study the movement of genes across natural populations. Work in Chapter III uses RAD-Seq to assess admixture between wild westslope cutthroat trout and introduced rainbows. This work was completed in collaboration with Hohenlohe PA, Day MD, Amish SJ, Miller MR, Boyer MC, Muhlfeld CC, Allendorf FW, Johnson EA and Luikart G. Work done for this dissertation also presents an adaptation of RAD-Seq to study restriction enzymes themselves. Chapter IV details the use of sequenced genomes to quantify restriction site preferences over complex substrates. This work was performed in collaboration with Quimby A, Zhu Z, and Johnson EA.

High-throughput sequencing experiments have been instrumental in a new era of gene expression studies[14,15]. RNA-Seq, the common term for inferring relative transcript abundance based on sequence counts, has become a common tool for probing transcriptomes in many systems[16,17]. In addition to providing a platform for quantifying expression across the transcriptome, RNA-Seq has revealed an exciting level of complexity by identifying novel gene isoforms[18,19]. The ability to use sequencing to identify exon junctions has uncovered that the very product expressed from a gene locus may be fundamentally different across cell states. Deep views of the RNA landscape have also identified transcript classes outside the central dogma. These non-coding

RNAs have a diverse set of functions and geneses[20,21,22] and reflect pervasive genomic transcription.

In addition to RNA-Seq, a series of other gene-expression methods have been developed for use on massively parallel sequencing machines. These techniques exploit the counting power of sequencers by linking biological events to sequencing events. ChIP-Seq involves the enrichment and sequencing of DNA bound to transcription factors in order to infer their binding sites[15] and FAIRE-Seq[23] is used to quantify chromatin accessibility. More recently, massively parallel reporter assays (MPRAs) have emerged which directly measure DNA regulatory activity[24,25]. These assays involve libraries of potential DNA transcriptional enhancers engineered such that their activity may be quantified in parallel by deep sequencing of the RNAs they produce. Combined with computational and statistical methods for analyzing the multidimensional data generated by these experiments, massively parallel sequencing experiments have brought unprecedented depth and discovery to gene regulation studies. We use new variations on these techniques to explore hypoxic gene regulation in chapter V. This work was done in collaboration with Preston JL, Randel MA, and Johnson EA.

CHAPTER II

RAPID MAPPING OF MUTATIONS IN *C. ELEGANS* BY RAD-SEQ

This work was published in Volume 189 of the journal of *Genetics* in 2011 in collaboration with O'Rourke SM, Yochem J, Connolly AA, Price MH, Carter L, Lowry JB, Turnbull DW, Stiffler N, Miller MR, Johnson EA, and Bowerman B. I contributed the molecular work and computational analysis pertaining to RAD-Seq and associated writing.

INTRODUCTION

Determining mutant gene identity is a key step for understanding gene function in forward genetic screens following mutagenesis and phenotype-based mutant isolation. In some organisms such as fungi and bacteria, a recessive mutant allele can be complemented with a plasmid-borne wild-type gene to establish gene identification. In organisms that lack robust DNA transformation methods, mapping with visible or selected single nucleotide polymorphism (SNP) markers to progressively finer genomic intervals is the traditional route to ascertain identity of the mutant gene. Now whole genome sequencing (WGS) methods can significantly reduce the time required to identify the causal mutation. For example, WGS can simply be used to determine all of the sequence alterations present in a mutant strain[1-5]. However, some mapping data are still required to differentiate the background mutational load from the causal mutation. More recently, WGS has been performed on outcrossed mutant progeny to combine mapping and sequencing for pinpointing the position of the causal mutation[6,7].

While resequencing a genome to identify mutant alleles is being used more frequently, in some cases it is more efficient to sequence only a portion of a genome. For example, sequencing of a single chromosome, a defined genomic interval, exonic sequences, or a single locus can be more cost effective when there is evidence that a mutation resides within a specific genome feature. There have been several throughput-enhancing advances in capturing targeted regions of a genome using DNA annealing since the first reported use of this methodology whereby individual microarray spots were physically scraped from the substrate[8-10]. For example, genomic DNA can be annealed to microarrays printed with oligonucleotides covering the region to be targeted, washed, and then eluted for sequencing[11-13]. Alternatively, oligonucleotides can be used to capture homologous genomic DNA in solution[14]. While these approaches are extremely high throughput, they also can be prohibitively expensive.

We have developed two Illumina-based sequencing methods in *Caenorhabditis elegans* that offer an alternative pipeline for mutation detection. First, we have performed restriction site associated DNA (RAD) polymorphism mapping to position the causal mutation to a relatively small region of the genome. Second, we have used genome interval pull-down sequencing (GIPS) to sequence a defined genomic interval. Genome intervals are captured by annealing sheared genomic DNA to sheared fosmids containing wild-type *C. elegans* DNA, eliminating the need for customized microarray or oligonucleotide production. Because multiple RAD mapping and genome interval sequencing samples can be combined in a single Illumina lane, it is possible to positionally clone and identify the mutant loci rapidly and cost effectively without performing WGS.

MATERIALS AND METHODS

C. elegans strains and culture

Strains were grown under standard laboratory conditions[15]. The temperature-sensitive (ts) mutants were maintained in a 15 °C incubator and shifted to a 26 °C incubator to perform temperature upshifts for determining embryonic lethality.

Genetic crosses for RAD mapping

To map or1167ts, we crossed the polymorphic *C. elegans* strain CB4856 into the original mutagenized background [or1167ts/or1167ts; lin-2(e1309)/lin-2(e1309)]. After self-fertilization of the heterozygous F1 outcross, we pooled 200 of the 1/16 of the F2 progeny that were again or1167ts/ or1167ts; lin-2(e1309)/lin-2(e1309), taking advantage of the lin-2 egg-laying defect to identify with a stereomicroscope within F1 self-progeny or1167ts/or1167ts; lin-2 (e1309)/lin-2(e1309) F2's filling up with dead embryos [avoiding laboriously singling out hundreds of F2's to look for production of dead embryos by egg-laying lin-2(+/+ or +/e1309) F2 progeny]. Similarly, for mapping unc-13, we also performed a cross to CB4856 but selected 200 Unc F2 progeny for the genomic DNA preparations. For mapping or1089ts, we crossed the original mutant to CB4856 males and isolated 800 F2 hermaphrodites that were tested for embryonic lethality. Approximately 200 homozygous or1089ts animals were recovered and used for the RAD mapping procedure.

Illumina sequencing for RAD mapping

Genomic DNA was isolated from pools of 200 homozygous unc-13 F2's, and 200 or1054ts; lin-2(e1309) F2's as well as the N2 and CB4856 parental strains using the

Qiagen DNeasy kit. A total of 150 ng of each sample was digested with EcoRI and processed into barcoded RAD libraries as previously described[17] with the minor modification of using the paired end P2 adapter[18]. Briefly, each sample was individually digested with EcoRI, and a P1 adapter (with a 4-bp barcode; see below) was ligated to the overhangs. After this step multiple samples were multiplexed. Next, the DNA was sheared and gel extracted to obtain 400-bp fragments and the Illumina P2 adapter was ligated. Samples were then run on the Illumina flow cell. The RAD library from the mutant pool was sequenced at 30X coverage in an Illumina Genome Analyzer IIx machine. With SNPs present at about every 1000 bp in the polymorphic CB4856 strain, and sequencing reads of 75 bp from each EcoRI site, we anticipated detecting a SNP near 1 in 10 EcoRI sites, or about one every 50,000 bp, which was close to the observed value of one SNP in 64,000 bp achieved, on average. The RAD sequences were aligned to the reference Bristol N2 genome using the Bowtie software package[19]. The Bowtie output was then exported to SAMtools[20] and converted into BAM format. We then produced a pileup file, to which we applied the samtools.pl script **"**varFilter**"** command (using default options) to identify SNPs. The varFilter results were then saved as a tab delimited file for use with graphing software (Microsoft Excel and Adobe Illustrator). As an alternative method for identifying N2/CB4856 SNPs, one could use the MAQGene program[21], which may be more accessible to non-bioinformaticians[6].

Illumina sequencing of genomic intervals

To pull down intervals of genomic DNA to which or195ts, semidominant or600(sd),ts, and or683ts were mapped, we used magnetic bead pull-downs. A total of 5 ug of genomic DNA was purified from each mutant strain using a DNeasy Blood and

Tissue kit (Qiagen) and sheared to an average size of 500 bp by sonication in a Bioruptor (Diagenode). The ends of the sheared DNA fragments were blunted using a QuickBlunt kit (New England Biolabs) and the fragments purified with a PCR purification kit (Qiagen). A-overhangs were added to the genomic fragments by incubation of the purified, blunted DNA with 150 units of Klenow DNA polymerase exo- (New England Biolabs) and dATP at 37 °C for 30 min. The modified fragments were purified with a mini-elute PCR purification kit (Qiagen). A total of 7 ul of 1 uM modified Illumina sequencing adapters were ligated to the sheared genomic fragments at 16 °C for 2 hr using 2000 units of T4 DNA ligase (New England Biolabs). [Top strand: 59 ACACTCTTTCCCTACACGACGCTCTTCCGATCxxxx*T 39; bottom strand: 59 phosphate xxxxGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG 39 (where x indicates the barcode bases and *, phosphorotioate bond.) The ligation reaction was size separated by agarose gel electrophoresis, and fragments between 150 and 500 bp in size were purified from the gel using a Gel Extraction kit (Qiagen). The purified ligation products were PCR amplified using Phusion high fidelity DNA polymerase (New England Biolabs) and the Illumina amplification primers 5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC GATCT 3' and 5' CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCT GCTGAACCGCTCTTCCGATCT 3'. The following cycling conditions were used for PCR: 98 °C for 2 min, 15 cycles of 98 °C for 10 sec, 65 °C for 30 sec, and 72 °C for 15 sec. Following amplification, samples were size separated by agarose gel electrophoresis, and fragments between 150 and 500 bp were purified with a Gel Extraction kit (Qiagen).

Biotinylated probe preparation

DNA preps homologous to the targeted genomic interval were prepared from genomic fosmids, using a nearly genome-wide fosmid library for *C. elegans* that was developed by the Genome Sciences Centre in Vancouver, BC, Canada. A total of 100 ng of the fosmid DNA mixtures were combined with 20 ul of 2.5 uM random octamer solution (Life Technologies) and heated to 100 °C for 5 min. The mixture was rapidly cooled in an ice/water bath, following which 5 ul of biotin dNTP mixture [1 mM biotin-14- dCTP, 1 mM dCTP, 2 mM dATP, 2 mM dGTP, and 2 mM dGTP, in 10 mM Tris-HCl (pH 7.5), 1 mM Na2 EDTA] (Life Technologies) was added, along with 1 ul of Klenow fragment DNA polymerase (Life Technologies) and ultrapure water to bring the reaction volume to 50 ul. The reaction was then incubated at 37 °C for 1 hr, following which, the products were size separated on an agarose gel and the predominant 100-bp product was purified with a Gel Extraction kit (Qiagen).

Streptavidin bead preparation

A total of 50 ul of M270 streptavidin Dynabeads (Life Technologies) were washed three times with 100 ml of 6x SSC and resuspended in 100 ul of bead block buffer [2% I-Block (Tropix), 0.5% SDS, 1X PBS]. Beads were incubated at room temperature for 30 min with occasional mixing and were then magnetically captured and washed three times with 6x SSC.

Hybridization, immobilization, elution, and sequencing

A total of 5 mg of adapted, purified genomic DNA was combined with 150 ng of purified biotinylated probe in 300 ul of hybridization buffer [54% formamide, 1· SSC, 1% SDS, 5.4X Denhardt's solution (Sigma), 1 mg/ml Salmon sperm DNA (Life

Technologies)]. The mixture was heated to 100 °C for 2 min and then transferred to a 42 °C incubator, where it was incubated with mixing overnight. Following overnight incubation, biotinylated probe/genomic DNA fragment hybrids were immobilized by binding to prepared blocked and washed streptavidin beads by combining the hybridization mixture (300 ul) with the bead/SSC mixture (100 ul), and incubating at room temperature for 15 min with occasional mixing. Beads were then magnetically captured, and washed three times with wash solution 1 (1X SSC, 0.15% SDS), three times with wash solution 2 (0.2X SSC), and three times with wash solution 3 (0.05X SSC). After the final wash step, the beads were resuspended in 200 ml of ultrapure water, heated to 100 °C for 2 min, and quickly magnetically captured. The supernatant was carefully collected and concentrated to a volume of 20 ul in a Speedvac concentrator (Savant). A total of 10 ul of the concentrated supernatant was then used as template for a PCR reaction utilizing Illumina amplification primers and Phusion high fidelity DNA polymerase (New England Biolabs) (2 min 98 °C 24 cycles of 98 °C for 10 sec, 65 °C for 30 sec, and 72 °C for 15 sec). The PCR products were purified with a PCR cleanup kit (Qiagen), quantified, and submitted for Illumina sequencing on an Illumina Genome Analyzer II.

RESULTS

RAD mapping of *C. elegans* mutations

To rapidly map *C. elegans* mutations, we have used an Illumina sequencing-based genome-wide single nucleotide polymorphism mapping procedure called RAD polymorphism mapping[17,22,23]. RAD markers are SNPs adjacent to restriction enzyme

recognition sequences in the genomes of divergent strains. In our case, we used the N2 background (isolated in Bristol, UK) to isolate mutants and subsequently crossed them to the polymorphic Hawaiian CB4856 strain for mapping. The N2 and CB4856 genome sequences have diverged substantially but their hybrid progeny are fertile. On average, there is a SNP approximately every 1 kb, allowing physical mapping using a large number of markers.

To experimentally identify RAD tags, we crossed wild-type N2 hermaphrodites to CB4856 males. F1 hybrid progeny were isolated and genomic DNA was digested with EcoRI. After adapter ligation and selective amplification of the RAD tags (Figure 1), Illumina sequencing was performed with an Illumina Genome Analyzer IIx system.

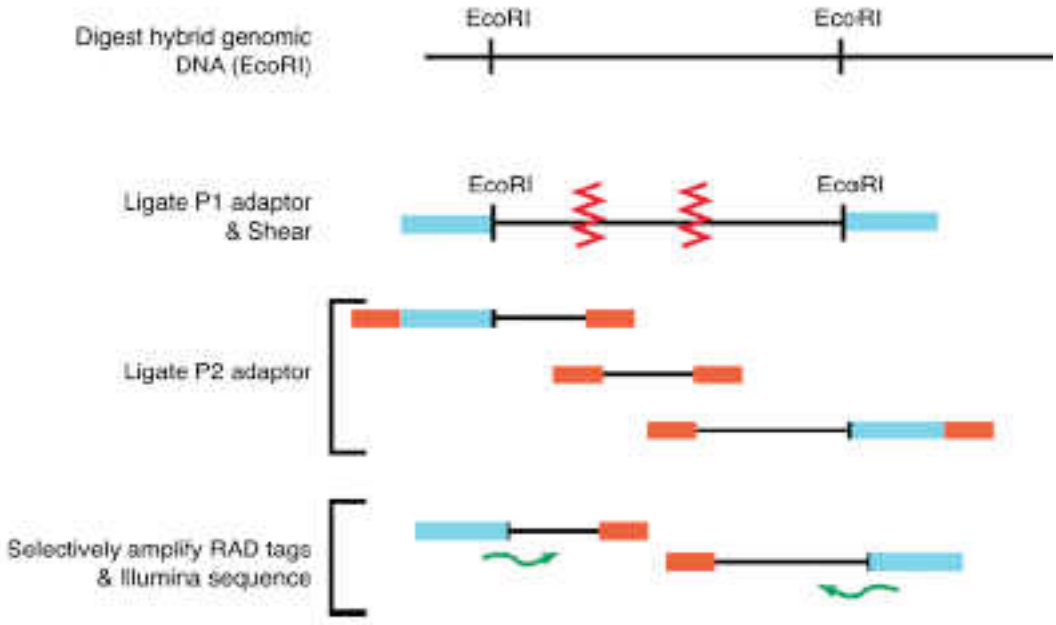

Figure 1. Restriction site-associated DNA (RAD) mapping schematic. Genomic DNA was isolated from 200 pooled F2 progeny in crosses between the N2 mutants and the polymorphic CB4856 (Hawaiian) strain. Genomic DNA was digested with EcoRI and P1 Illumina adapters were ligated to the fragments. This DNA was mechanically sheared and Illumina P2 adapters were ligated to the fragment ends. Next, RAD tags were selectively amplified and sequenced from the Illumina sequencing primer site encoded on the P2 adapter on an Illumina Genome Analyzer IIx machine.

Selective amplification was carried out by using a ‘‘Y’’ adapter for the P2 adapter, which prevents fragments that lack a P1 adapter from being amplified after first round synthesis initiated from the P1 site [17,24]. We detected 3,462 SNPs with an average distance between them being 29 kb. Most SNPs (95%) were separated from an adjacent SNP by 100 kb (Figure 2). The largest distance separating adjacent SNPs occurred near the center of chromosome V (515 kb). Most of the SNPs we identified could be predicted in silico from the sequence of the CB4856 strain (data not shown). Because the sequencing is

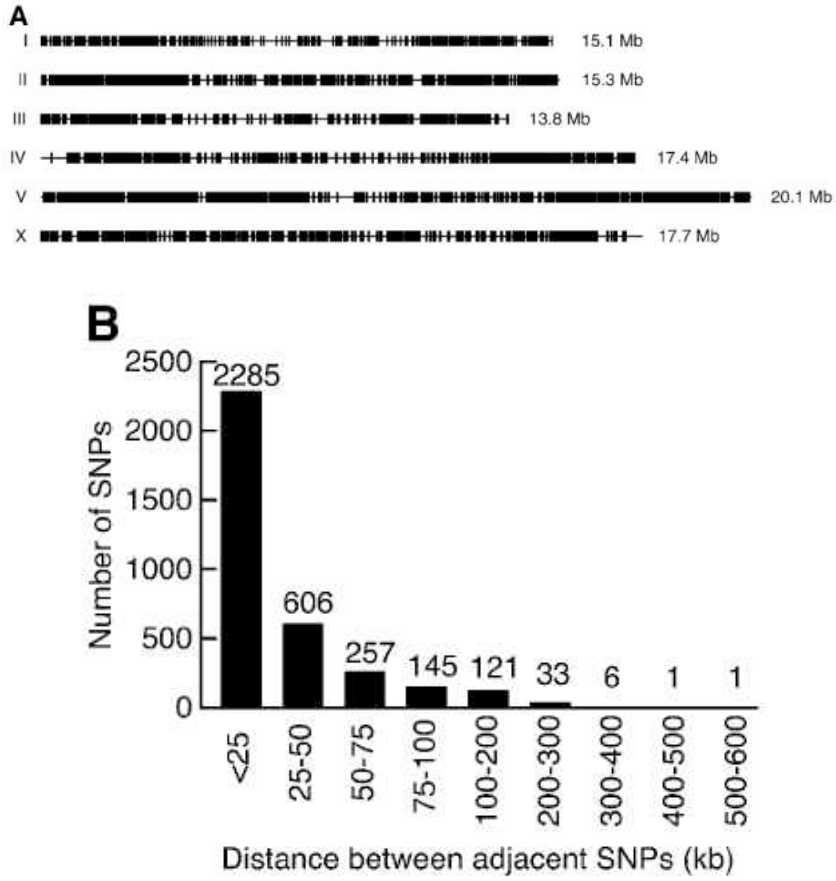

Figure 2. EcoRI-associated RAD tag locations. (A) RAD mapping results from an N2/CB4856 cross. Vertical lines represent EcoRI-associated RAD markers on each of the C. elegans chromosomes. The total chromosome sizes are listed on the right. For a list of the RAD markers, sequences, and positions, see File S1. (B) Distance between adjacent RAD markers. The plot represents the number of RAD marker pairs at the indicated distances, in kilobase pairs.

done with purified RAD tags instead of total genomic DNA, multiple samples can be multiplexed on a single lane in an Illumina sequencer, with each sample containing unique barcodes or subsequent sequence data deconvolution. The barcodes used for RAD mapping are 6-bp sequences added to the P1 adapter primer. In one test, we used one Illumina Genome Analyzer IIx lane to process 13 RAD mapping crosses. We used single-end sequencing with 80-bp reads, to achieve 50 million reads yielding 40· coverage. We tested the applicability of RAD mapping coupled with Illumina sequencing using three different approaches.

First, we mapped a known mutant, unc-13(e450). We crossed the unc-13 mutant to the polymorphic *C. elegans* strain CB4856 and pooled 200 F2 progeny that were homozygous for the unc-13 mutation. We chose 200 recombinants as a goal because it afforded a relatively large number of independent recombination events, although it is possible that using fewer recombinant F2's would also yield sufficient resolution. After producing a RAD library from the F2 genomic DNA sample, we performed sequencing on the Illumina machine to detect SNPs across the genome (see Materials and Methods). We used graphing software (Microsoft Excel and Adobe Illustrator) to plot the ratio of CB4856/Bristol SNPs across the *C. elegans* genomic sequence, indicating the fraction of samples for any one SNP that correspond to the polymorphic CB4856 sequence (Figure 3A). In this test, we identified 683 RAD tags throughout the genome. The ratio of CB4856 to N2 SNPs was 0.5 across the genome, except for chromosome I, where a large trough was present. The center of the trough on chromosome I is within 800 kb of the known location of unc-13 (Figure 3B). We conclude that Illumina-based RAD mapping can quickly provide the approximate physical position of mutant loci.

We similarly applied RAD mapping to or1089ts, a mutant of unknown molecular identity with severe defects in mitotic spindle assembly in one-cell–stage embryos (Figure 4A). In this case, we identified 3134 RAD markers in the F2 RAD library. Chromosome I was highly enriched for N2 DNA (Figure 4B), with the trough of N2 DNA centered at 9.25 Mb on chromosome I (Figure 4C), positioning the or1089ts mutation near the center of chromosome I. In a 1-Mb region centered on the trough, we

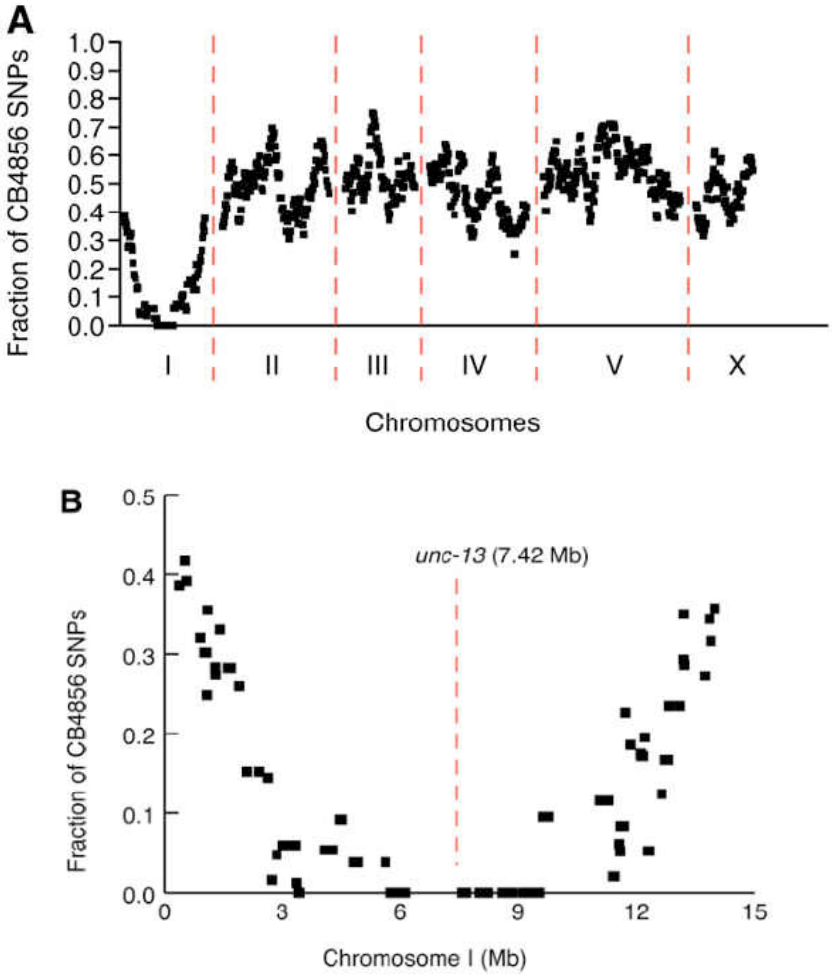

Figure 3. RAD mapping results for unc-13. (A) Genome-wide RAD mapping results for unc-13 after crossing it to CB4856. A total of 683 SNPs across the genome in F2 progeny were detected and the ratio of the CB4856 SNPs plotted along the chromosomes. The vertical axis represents the percentage of CB4856 SNPs in the F2 population. (B) Magnification of chromosome I. The trough on chromosome I correlates with the known location of unc-13.

found one candidate gene in online databases, spd-2, that when reduced in function using RNAi, results in defects that closely resemble the or1089ts mutant phenotype[25,26]. The center of the reduced CB4856 ratio is 276 kb to the left of the known location of spd-2 (Figure 4C). We Sanger sequenced the spd-2 gene in genomic DNA from or1129ts mutants after amplifying the locus using genespecific primers. We found a single

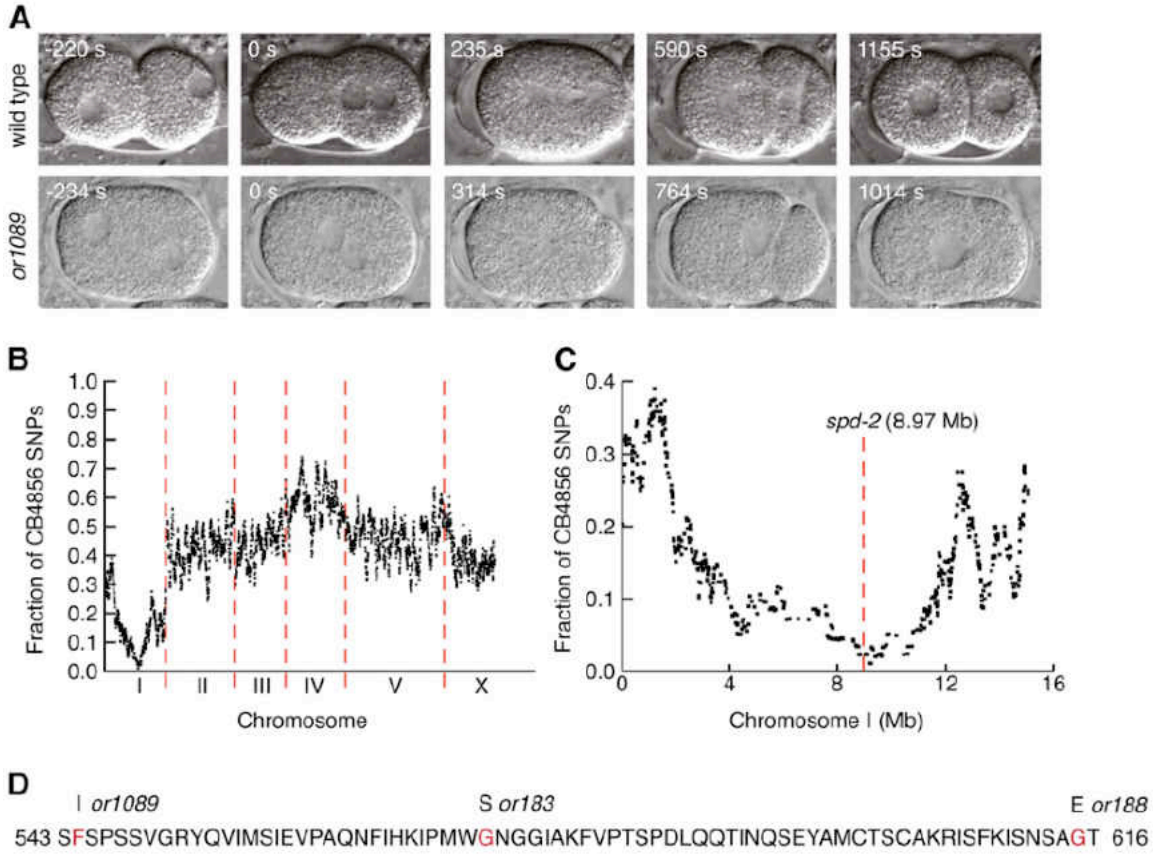

Figure 4. Phenotype and RAD mapping of spd-2(or1089ts). (A) Defective mitotic spindle formation and cytokinesis failure in an early or1089ts embryo produced from a worm shifted to 26_ for 6 hr (lower) compared to a wild-type embryo (upper). (B) Genome-wide RAD mapping of or1089ts. A reduction of CB4856 DNA on chromosome I results from the selection of the mutant homozygotes. (C) The center of the trough is near 9 Mb and is positioned near the causal mutation (see text). (D) DNA sequence analysis of or1089ts identifies amutation in spd-2 open reading frame (GenBank: AY340594.1). The mutation, causing a phenylalanine-to-isoleucine change in codon 544, is shown relative to the changes of or183 and or188, two known temperature-sensitive alleles of spd-2

nucleotide mutation that changes a phenylalanine codon to an isoleucine codon at amino acid 544 (Figure 4D). We also found that or1089ts failed to complement spd-2(or293ts) (data not shown). We conclude that or1089ts is a new spd-2 allele.

We have screened for temperature-sensitive, embryonic lethal *C. elegans* mutants using an egg laying-defective (Egl) lin-2(e1309) mutant background that enables one to screen mutagenized populations for animals that produce inviable embryos without singling out individual worms[16,26,27,29]. To take further advantage of this Egl background for RAD mapping, we crossed or1167ts; lin-2(e1309) hermaphrodites to CB4856 males and obtained F2 animals that were mutant for both lin-2(e1309) and or1167ts. Instead of testing embryonic lethality of 800 F2 animals individually on plates, as was done for spd-2(or1089ts), we were able to isolate 200 or1167ts; lin-2(e1309) animals that accumulated mostly dead embryos from a mixed F2 population after shifting them to the restrictive temperature as L4 larvae. After preparing and sequencing the RAD library, we identified 3400 RAD tags. The ratios of CB4856 to N2 DNA were plotted and we found two troughs that correspond to an enrichment of N2 DNA (Figure 5). The trough on the X chromosome is 550 kb from the known location of lin-2, while the trough on the right arm of chromosome IV presumably correlates with the location of or1167ts. The fact that the lin-2 trough does not reach zero likely relates to inadvertently picking some worms that were in fact not lin-2, as even wild-type worms sometimes hold their embryos and can appear Egl. We found one candidate gene, sas-6, positioned 250 kb left of the center of the chromosome IV trough. Like animals depleted for sas-6, or1167ts animals shifted to the restrictive temperature as L4 larvae produce embryos that appear to assemble monopolar mitotic spindles in early embryonic cells (not shown). We sequenced the sas-6

locus with the Sanger method after amplification of the region by PCR. We found a single missense mutation that changes an aspartic acid to a valine in the ninth amino acid of SAS-6 (Figure 5D). As or1167ts also failed to complement a known sas-6 mutant (not shown), we conclude that or1167ts is a sas-6 allele. This example demonstrates that RAD mapping can be easily applied in the context of the lin-2(e1309) marker to minimize the effort required to isolate F2 animals that are homozygous for embryonic lethal mutations. We continue to explore the use of RAD methodologies to rapidly map temperature-

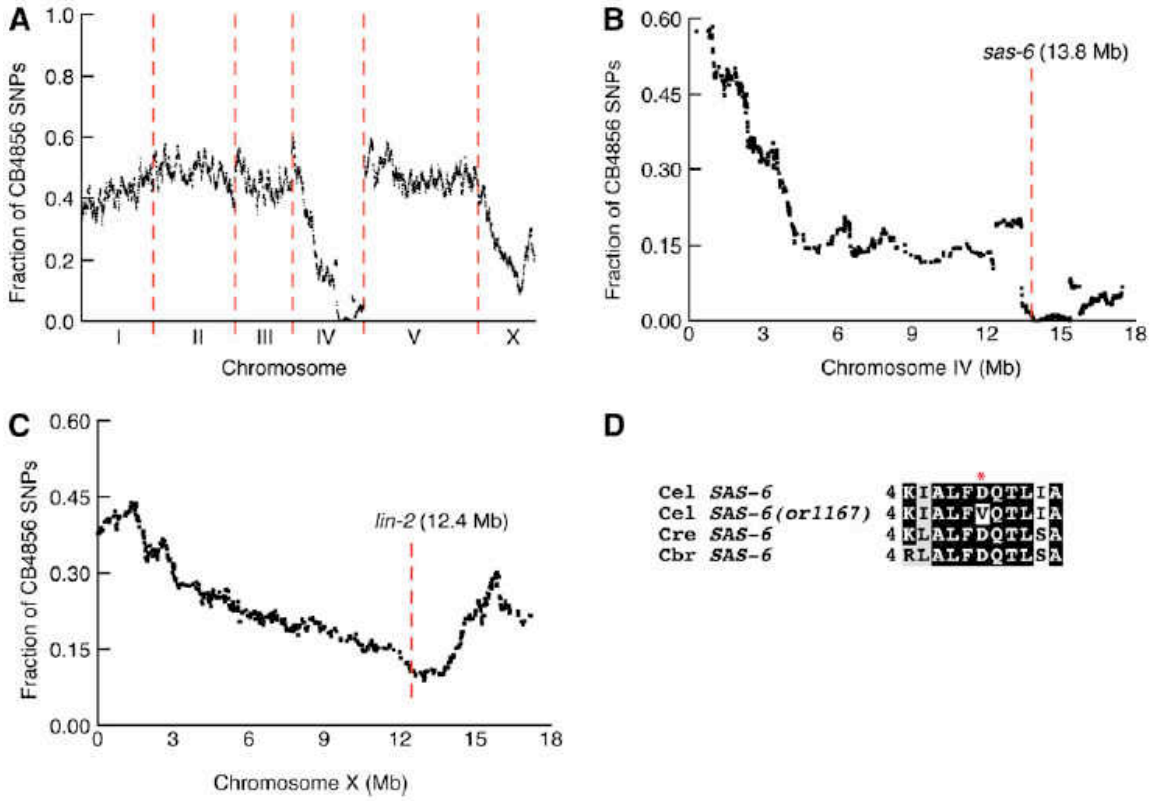

Figure 5. RAD mapping results for a sas-6(or1167ts); lin-2 double mutant. or1167ts; lin-2 mutants were crossed to CB4856 males and F2 progeny, homozygous for both the or1167ts embryonic lethal mutation and the lin-2 mutation, were isolated. (A) Genomewide RAD mapping results show enrichment for N2 DNA on chromosomes IV and X. (B) The sas-6 gene is located near the chromosome IV trough. (C) The trough on the X chromosome lies near the known position of the lin-2 locus. (D) The sas-6 locus contains a missense mutation that changes an aspartic acid codon to a valine (NCBI: NP_502660.1). An alignment was performed with the C. elegans (Cel) wild-type SAS-6, C. elegans SAS-6(or1167ts), C. remanei (Cre) SAS-6, and C. briggsae (Cbr) SAS-6.

sensitive embryonic lethal *C. elegans* mutants and note that this approach can be used to map virtually any locus that can be assayed in N2/CB4856 F2 animals.

Illumina-based GIPS

To quickly and more cost-effectively identify causal mutations in mutant strains, we have also applied Illumina DNA sequencing to defined genomic intervals, rather than sequencing entire mutant genomes (as has been done to identify some mutant loci in *C. elegans*[1,7]). Sequencing an entire genome involves more cost than our procedure because we multiplexed multiple barcoded sequencing experiments on a single Illumina Genome Analyzer IIx flow cell. Briefly, we used wild-type genomic DNA from defined genomic intervals, linked to magnetic beads, to partially purify regions of mutant genomic DNA (see Figure 6A and Materials and Methods). We first tested the feasibility of using interval pull-downs and Illumina sequencing by resequencing a previously identified mutation present in the dhc-1 locus of dhc-1(or195ts) mutants worms[30]. We then identified the mutations responsible for conditional lethality in two previously reported mutants [tbb-2(or600sd,ts) and plk-1 (or683ts)[29]], after sequencing genomic intervals of 1.8 and 1.3 Mb, respectively (Figure 7).

To test this methodology, we first selected a fosmid that includes the entire dhc-1 locus, available from Source Bio- Science (http://lifesciences.sourcebioscience.com). We purified the fosmid, sheared it, and linked the fragments to biotinylated beads as described in Materials and Methods. After using these beads to isolate sheared dhc-1(or195ts) genomic DNA, the mutant genomic DNA was eluted and prepared for Illumina sequencing (see Materials and Methods). We aligned 66,853 30-base reads to the 30.1 kb dhc-1– containing fosmid on chromosome I and the average coverage for

each position in the fosmid was 66·. For comparison, we also identified 570,311 reads that could be aligned to the rest of the genome, yielding an average read coverage of 0.17· for each nucleotide position. Therefore, we achieved a 388-fold enrichment for reads in the targeted region using our interval pull-down sequencing method. We identified the previously sequenced C-to-T dhc-1(or195ts) mutation in a total of 66 reads, and no other mutations were detected in the dhc-1 locus (Figure 7A). We conclude that GIPS can readily identify the mutations present in relatively large regions of the genome.

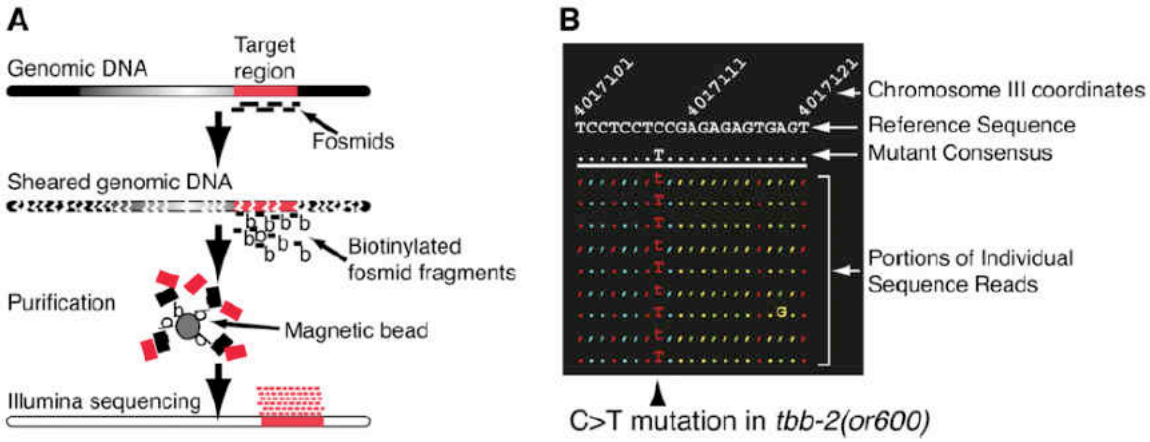

Figure 6. Genome interval pull-down sequencing (GIPS) using the Illumina platform. (A) Schematic overview of the interval pull-down sequencing method. First, fosmids of wild-type DNA covering a region of the genome are purified, sheared, and ligated to biotinylated adapters. Next, mutant genomic DNA is sheared and annealed to the biotinylated fosmids. After purification and release of the mutant DNA using magnetic beads, the fragments are subjected to sequencing on an Illumina machine. Finally, the reads are assembled onto the genome scaffold and polymorphisms are identified. (B) Example output of mutant genome assembly. Shown is a small region of the tbb-2 locus with portions of reads aligned beneath the reference sequence. In this case, each read shows that a cytosine in wild type has been changed to a thymidine in tbb-2(or600sd,ts). Also shown is one apparent sequencing error where an A . G change was called in one of the reads. Nucleotides are color coded, uppercase letters, and periods in the sequence reads represent identity with the reference sequence, while reads containing lowercase letters and commas represent sequence data obtained from the reverse complement strand.

Next, we used GIPS to identify the mutation responsible for the early embryonic cell division defects caused by the semidominant, temperature-sensitive mutation

or600sd,ts. We defined the or600sd,ts interval using standard mapping crosses with visible morphological and behavioral markers. We localized the mutation to chromosome III between positions 3,618,381 and 5,447,436. We used the WormBase genome browser to identify a minimal tiling path using genomic DNA from fosmid clones available from Source BioScience (http://lifesciences.sourcebioscience.com). We identified 65 fosmids that spanned the region with 7 gaps that totaled ,45 kb (2.5% of the region). We purified, sheared, and linked the fosmid fragments to biotinylated beads. After using these beads to isolate sheared or600sd,ts genomic DNA, the mutant genomic DNA was eluted and prepared for Illumina sequencing (see Materials and Methods). We found 1,596,403 48-base reads that could be aligned to the region, giving an average coverage for each nucleotide in the interval of 42·. The total number of reads corresponding to the *C. elegans* genome was 6,034,221.

We used software (SAMtools) to output a text file that lists the mutations identified in the interval. One can also view the sequence reads aligned to the reference wild-type genome sequence with SAMtools (Figure 6B). There were 45 mutations in the 1.8-Mb or600sd,ts interval (Figure 7B). We found 14 extragenic changes, 23 intronic changes, one mutation in a pseudogene, two mutations that cause the transcript to go out of frame, three potential annotation errors, and two missense mutations. As 90% of our identified temperature-sensitive mutations are caused by missense mutations[29] (the remaining are due to premature stop codons, small deletions, and mutations in splice-site boundaries), we narrowed our analysis of mutations in exon and intron splice sites. Three mutations that appeared to cause exonic changes are likely not the causal mutation because they were present in wildtype DNA; expressed sequence tags show the same

mutations, perhaps indicating errors in the reference sequence. None of the intronic

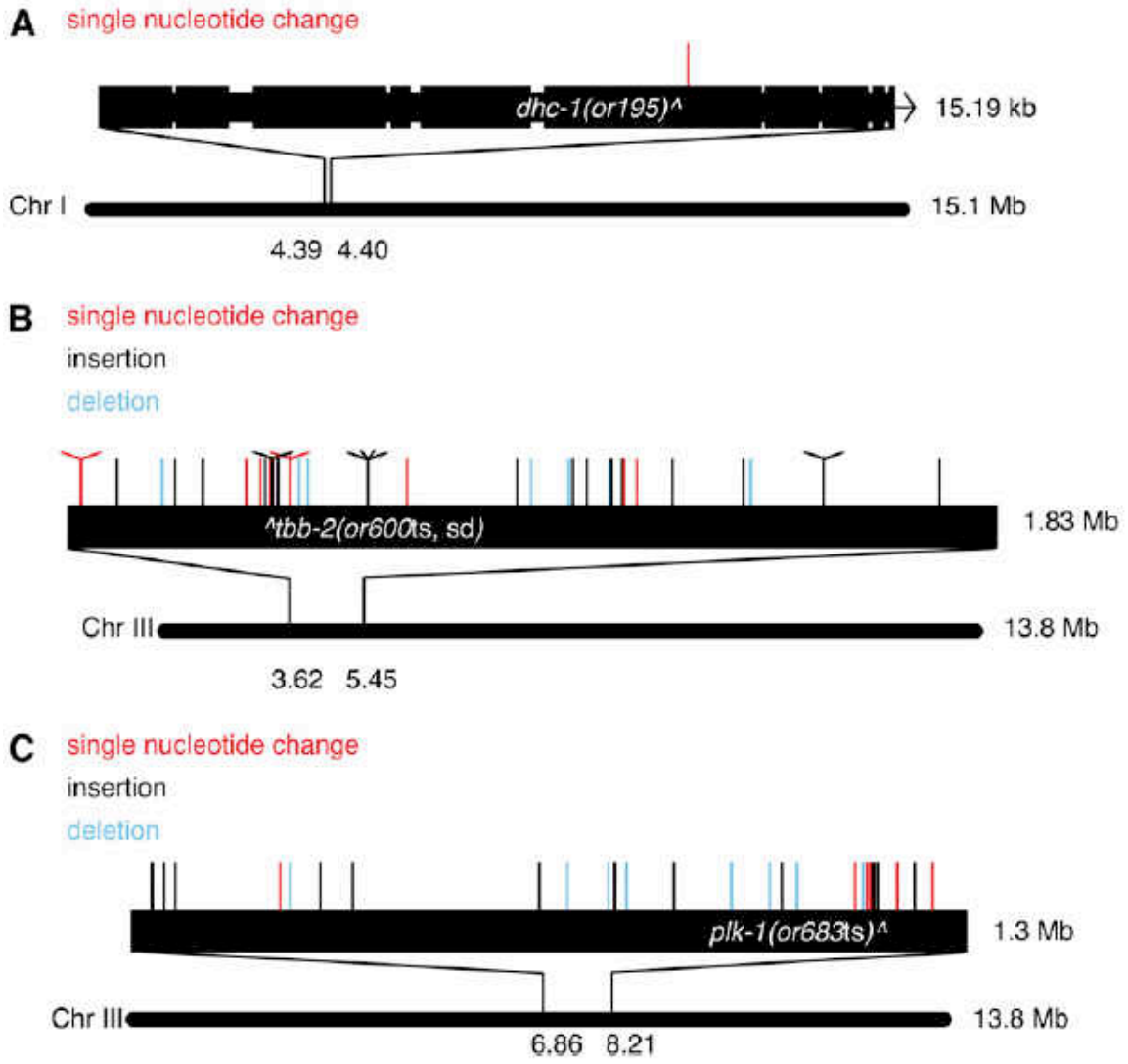

Figure 7. Analysis of the genome interval pull-down sequencing. The protocol shown was applied to three mutants. The positions of nucleotide alterations vs. the reference sequence is shown in graphical form. Each line represents a single change, diagonal lines attached to the top of the vertical lines represent multiple changes located very close together. Lines are color coded to show various types of mutations. The ^ in the black bars point to the causative mutations. (A) Sequencing results for the 15 kb dhc-1(or195ts) locus purified using a single genomic fosmid clone. The previously identified missense mutation was found. (B) Sequencing results for a 1.82-Mb region on chromosome III in the or600sd,ts genome. We identified 45 total mutations in the region including the missense mutation in tbb-2. (C) Sequencing results for a 1.3-Mb region on chromosome III in the or683ts genome. We identified 29 total mutations in the region including the missense mutation in plk-1.

changes occurred at intron boundaries and thus are unlikely to interfere with RNA splicing. The four remaining exonic mutations occur in the tbb-2, ras-2, his-70, and clec-154 genes. Single nucleotide deletion and insertion mutations in the his-70 and clec-154loci, respectively, encode proteins with altered C termini, while the ras- 2 and tbb-2 mutations are missense. Of these four genes, only RNAi that targets the tbb-2 locus phenocopies the or600sd,ts early embryonic phenotype [note that tbb-2(RNAi) also depletes the paralogous redundant tbb-1 gene product]. Depletion of the other three genes by RNAi does not result in any lethal phenotypes (WormBase). On the basis of sequence data, the embryonic phenotype and genetic interactions with a previously isolated tbb-2 allele, or362sd,ts[29], we conclude that or600sd,ts is a tbb-2 allele.

DISCUSSION

The utility of *C. elegans* as an animal model, in which one can readily isolate temperature-sensitive mutations in essential genes, and the power of next generation DNA sequencing for greatly reducing the time required to positionally clone mutant loci, now make it possible to much more rapidly isolate experimentally useful conditional mutations in essential genes. Our two new Illumina-based sequencing methods should allow for increased throughput when analyzing large numbers of mutants.

RAD polymorphism mapping has been used successfully to map genes in *Drosophila*[23], threespine stickleback fish[23], Neurospora[22], diamondback moths[31], barley[32] and now *C. elegans*. So long as a hybrid strain is available to generate F2 progeny, the methodology should be feasible in any organism. RAD mapping makes it possible to simultaneously and rapidly determine the approximate location of large numbers of

mutations isolated after mutagenesis. We have used EcoRI to cut the genomic DNA because it provides relatively good resolution between RAD markers (Figure 1). However, any other restriction enzyme could also be used, or multiple enzymes could be used to gain increased mapping resolution. RAD mapping provides an ideal procedure to identify mutant loci when the number of gene candidates is limited.

We explored the use of RAD mapping in three different contexts. First we have performed a proof-of-principle experiment where we mapped the position of unc-13, a known mutant (Figure 3). While the numbers of RAD markers obtained was relatively low in this example (683, presumably because we experienced some sample loss), the center of the trough on chromosome I is still within 800 kb of the known position of unc-13. In the second approach, we used RAD mapping to clone the or1089ts mutant. After picking 800 F2 animals from an or1089ts/CB4856 cross, we identified those that were homozygous for the or1089ts mutation. We found a substantial enrichment of the N2 DNA on the center of chromosome I that was positioned within 276 kb from the known position of the spd-2 locus. We performed Sanger DNA sequencing on PCR products derived from the spd-2 locus and identified one sequence alteration that causes a missense mutation (Figure 4D). Thus, RAD mapping can rapidly identify candidate genes that can be further investigated by sequencing candidate genes, GIPS, or complementation tests with existing mutants. We also tested the feasibility of performing RAD mapping on F2 animals that were doubly mutant for an embryonic lethal mutation and an egg-laying defective mutant, lin-2 (present in the original mutagenized strain). Since homozygous lin-2 animals hold their embryos, it was possible to more easily and rapidly select F2 progeny homozygous for the mutation being mapped from an or1167ts;

lin-2 and CB4856 cross (Figure 5). As expected, we found two regions of the genome that were enriched for N2 DNA: one corresponds to the lin-2 locus, while the second corresponds to the or1167ts mutation in the sas-6 locus. Finally, we are currently exploring the use of RAD mapping by crossing temperature- sensitive mutants to CB4856 males and allowing the progeny to reproduce at the nonpermissive temperature for many generations. This method may significantly reduce the labor in isolating the homozygosed F2 progeny and would show an exclusion of N2 DNA corresponding to the lethal locus.

In a second use of Illumina sequence technology, we have developed a genome interval pull-down method to sequence defined portions of the genome. By sequencing only intervals that contain the causal mutation, one can reduce the expenses associated with whole genome sequencing. We have successfully applied this technology to positionally clone two new mutants so far (Figure 7), as well as the control dhc-1(or195ts) mutant. We identified 45 sequence alterations in the 1.83 Mb or600sd,ts interval and 29 mutations in the 1.3 Mb or683ts interval vs. the WormBase reference sequence. Thus, if we had sequenced the entire genome of these mutants, we would have found many mutations. Therefore, it is clearly important to have some mapping data to narrow the search for the causal mutation, and RAD mapping fills this role well. As costs continue to come down, whole genome sequencing using either recombinant F2 animals[6] or backcrossed mutants[7], two techniques that simultaneously map and sequence mutations, may become more cost effective than the relatively more labor-intensive GIPS. Nevertheless, RAD mapping may continue to prove useful for analyzing large numbers of mutants, as many mutants can be identified by sequencing only candidate

genes in the vicinity or by complementation tests with previously identified alleles in the region. In fact, a large number of nonconditional mutants exist that can be used for performing complementation tests[33-36]. GIPS should also remain useful for sequencing candidate genes that are too large to easily amplify with PCR for Sanger sequencing.

As of August 2011, the cost to sequence the entire *C. elegans* genome at 30X coverage was about U.S.$600 on the HiSeq2000 platform. With this many reads, one could perform 50 RAD mapping experiments (at 30X coverage) or 50 GIPS procedures using 2-Mb pull-down regions. Both of these sequencing techniques can also be run on Illumina runs with other samples (with the use of barcoded adapters). Depending on the type of mutant being sequenced, using WGS strategies will be more straightforward. For example, if the mutant locus is resistant to RNAi, or if large-scale RNAi screens have not assayed the phenotype represented by the mutant, then WGS is likely the best approach. However, if the mutant phenotype is likely to be recapitulated by RNAi, such as early embryonic lethality (as we are studying), then RAD mapping will reveal a limited number of candidate genes. Sanger sequencing, complementation tests with existing mutants, or GIPS could then identify the causal mutation, although the time required to perform this two-step approach is longer than using a mapping/WGS approach. Thus, RAD mapping may be a viable alternative to WGS when large numbers of mutants are being investigated. Similarly, GIPS can be useful for sequencing loci that are known to be defective in many different mutants. For example, suppressor screens often identify many intragenic suppressor alleles[37] which, depending on size, can be very time consuming to Sanger sequence; yet performing WGS may be too expensive with many alleles to sequence. GIPS fills this gap and allows the simultaneous sequencing of many

different mutant loci on the same Illumina lane with the use of barcoded samples. In conclusion, we offer two new strategies for mutant identification in *C. elegans* that can fill roles not currently provided by WGS for certain applications.

BRIDGE TO CHAPTER III

Having discussed the use of RAD-Seq generated markers to identify mutant loci in recombinant lab animals, we turn towards the application of RAD-Seq in natural populations. The complexity reduction provided by only analyzing the consistent subset of genomic regions adjacent to restriction sites allows for many individuals to be assayed in a cost-effective manner. It also is a next-generation method applicable to non-model organisms providing resolution and throughout far beyond sanger-based molecular population genomics studies. In analyses such as that presented in chapter III, polymorphic loci can simultaneously be identified and used to infer population structure. Allelic differences between populations can be used to create phylogenies as well as witness evolutionary events. In the following chapter the level of admixture between wild westslope cutthroat trout and introduced rainbow trout is traced using RAD-Seq. Similar to the *C. elegans* study described in chapter II, abnormalities in the distribution of polymorphic RAD sequences are used to infer biological activity. In this case they are not indicative of lab-produced mutations, but of introgressing alleles conferring fitness advantages in natural populations.

CHAPTER III

GENOMIC PATTERNS OF INTROGRESSION IN CUTTHROAT TROUT

This work was published in volume 22 of the journal *Molecular* Ecology in 2013 in collaboration with Hohenlohe PA, Day MD, Amish SJ, Miller MR, Boyer MC, Muhlfeld CC, Allendorf FW, Johnson EA and Luikart G. I contributed the molecular work pertaining to RAD-Seq and associated writing.

INTRODUCTION

Hybridization between native and introduced taxa is an increasing concern for conservation and legal assessments of threatened species[1]. Hybridization can reduce fitness through outbreeding depression[2], cause genomic extinction[1] and destroy important genetic and ecological adaptations[3,4]. The loci most responsible for the genetic effects of hybridization may be outliers in their degree of introgression because of natural selection in admixed populations[5-8] ('super invasive alleles'). As a result, estimates of admixture averaged across loci at the individual or population level may miss important genetic factors in conservation and management of native taxa. Current high-throughput sequencing techniques now allow genome scans for invasive alleles in natural populations of nonmodel species.

Anthropogenic hybridization is especially widespread in freshwater fishes due to decades of fish translocations and hatchery supplementation of wild populations. Rainbow trout (RBT, *Onchorhynchus mykiss*) is the most widely translocated and problematic invasive fish worldwide[9]. RBT hybridize with cutthroat trout (*O. clarkii*),

including the subspecies westslope cutthroat trout (WCT, *O. c. lewisi*). WCT is the most widely distributed of 12 extant cutthroat subspecies, and hybridization is the leading threat to persistence of genetically pure WCT populations[10].

Management of WCT populations would benefit from detection of hybridization and introgression at low levels and from the ability to precisely estimate individual-level admixture proportion. Previous work has used micro satellites and other loci to assess levels of admixture from RBT into native WCT populations[2,11,12,13]. Muhlfeld et al.[13] found that levels of RBT admixture were negatively related to distance from the source of RBT hybridization (Abbot Creek; see Figure 1) and positively related to mean summer water temperature, suggesting potential for the existence of RBT alleles that are adaptive to warm water temperatures[14,15]. However, the low number of diagnostic markers available with microsatellites typically allows precise admixture estimates only at the population level, not at the individual or genome-scan level.

Single nucleotide polymorphisms (SNPs) are ideal markers for hybridization assessment and monitoring because hundreds of SNPs can be rapidly, reliably and cheaply genotyped using new genotyping platforms[16-19]. Much recent effort has been committed to assembling a set of diagnostic SNP loci for RBT and WCT[20-25].

A high density of markers across the genome promises individual-level estimates of admixture proportion, as well as detection of super invasive alleles. However, SNP discovery in salmonid fish is especially challenging due to a recent genome duplication event, making it difficult to distinguish true SNPs from fixed sequence differences between homeologous duplicate chromosomal regions[26-28] as well as more typical tandem-duplicated paralogous regions. One way to filter out both paralogs and homologs

is to gather more sequence data around candidate SNP markers to resolve between next-generation sequence reads that come from one locus vs. two different loci.

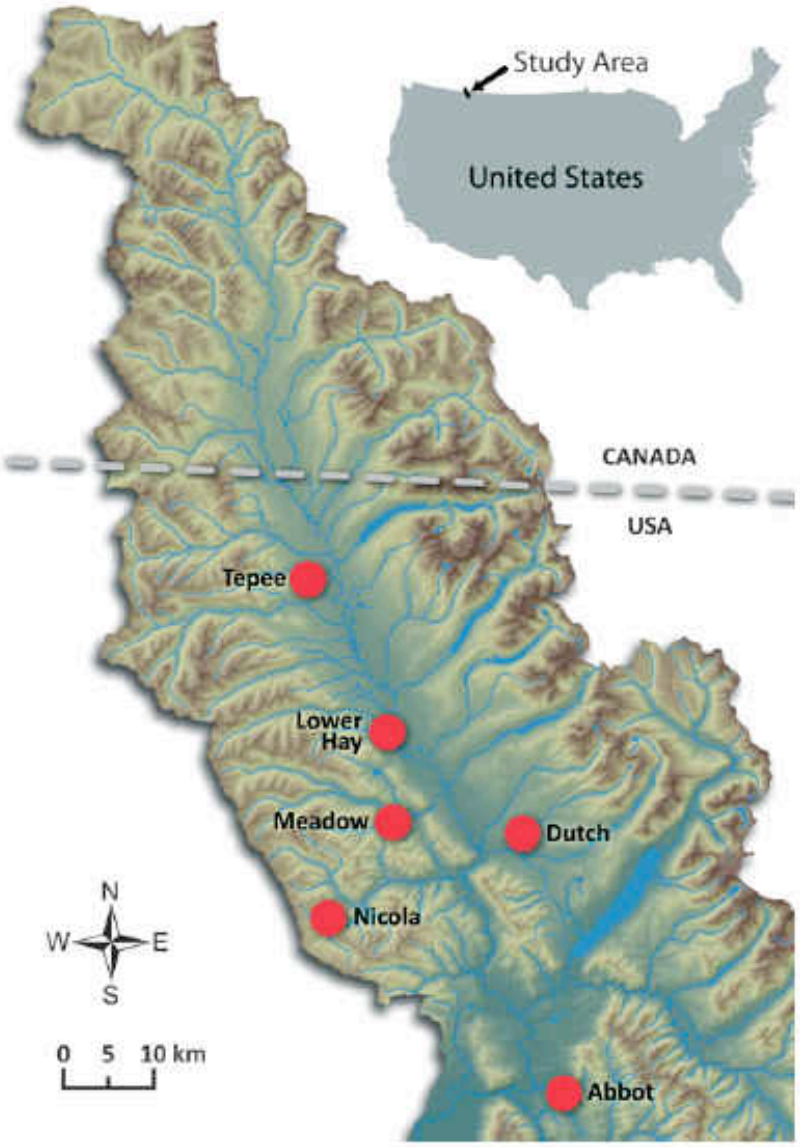

Figure 1. Map of the North Fork Flathead River study area, showing the five admixed westslope cutthroat trout populations examined here plus the initial source of introduced rainbow trout individuals (Abbot Creek; see Boyer et al. 2008; and Muhlfeld et al. 2009c for more information on these populations)

We previously used restriction-site-associated DNA (RAD) sequencing[29] to identify several thousand WCT diagnostic SNPs[30]. Those candidate diagnostic markers have shown a high rate of subsequent validation in microfluidic PCR-based genotyping assays[23]. However, primer design for those genotyping assays required >50 bp of flanking sequence on each side of each SNP, which we obtained from previously published sequence data, reducing the number of candidate markers for which assays could be designed[23]. In addition, our ability to distinguish duplicate sequence based on the flanking sequence was limited to the 54 bp single-end Illumina read length in that study. The approach we present here can be used to simultaneously identify and genotype SNP markers, as well as gather substantial flanking sequence, in a single RAD sequencing experiment. The amount of flanking sequence is more than sufficient for primer design and also allows better discrimination of paralogous loci.

Restriction-site-associated DNA sequencing is one of a family of genomic approaches that provide sequence data adjacent to restriction enzyme recognition sites[31]. The primary difference between RAD and related techniques is that RAD incorporates a random shearing step in library preparation. As a result, while the forward reads are anchored at the restriction site, the reverse reads produced by paired-end Illumina sequencing of RAD libraries are staggered over a local genomic region (of several hundred base pairs). These staggered paired-end reads can be assembled into a 'mini-contig', a continuous stretch of genomic sequence that is longer than each individual read and potentially up to 1 kb[32, 33, 34, 35]. Here, we designed our RAD libraries so that a substantial fraction of DNA fragments would produce overlapping paired-end reads, allowing assembly of contigs containing both the forward and reverse reads of each pair.

These ‘RAD contigs’ are anchored at one end by the restriction enzyme recognition site and contain several hundred base pairs of continuous genomic sequence data across dozens of individuals.

The goals of this study were to: (i) assemble a large set of RAD contigs from a sample of low-admixture WCT populations; (ii) provide flanking sequence for finer filtering of candidate diagnostic SNP markers between RBT and WCT; (iii) genotype filtered diagnostic SNPs across five WCT populations to assess the ability of RAD sequencing compared with microsatellites to provide precise individual-level estimates of admixture; and (iv) identify outlier loci exhibiting the signature of super invasive alleles.

MATERIALS AND METHODS

Study system

We focus on WCT populations in tributaries to the North Fork of the Flathead River in northwestern Montana. The North Fork Flathead River originates in Canada and forms the western border of Glacier National Park before joining the main-stem Flathead River, which flows into Flathead Lake. The presence of hybridization and RBT admixture was previously estimated in several populations using seven diagnostic microsatellite loci[12,13]. Here, we use five of these populations (Meadow, Nicola, Dutch, Lower Hay and Tepee) for which estimates of the mean population-level admixture based on microsatellite loci ranged from 1.3% to 13.0%. We chose populations without F1 hybrids as identified in previous studies with the goal of using later-generation admixed populations to detect specific loci with elevated levels of introgression. We used preserved DNA samples, collected from 18 to 22 individuals in each population during

2003 to 2004 for the study by Boyer et al.[12] to allow individual-level comparisons between SNP-based and microsatellite-based admixture estimates. We selected individuals across the range of admixture proportions previously estimated within each population.

RAD sequencing

We prepared RAD sequencing libraries for 97 samples from the five WCT populations described previously, following the previously published protocol.[36] The RAD protocol produces libraries of genomic fragments bounded on one end by a restriction enzyme cut site (therefore common across individuals), with the other end randomly sheared. Typically, fragments in RAD libraries are size selected simply to optimize the efficiency of the Illumina sequencing process. Here, we used the restriction enzyme SbfI and 6-nucleotide barcoded adaptors differing from each other by at least three nucleotides to identify individuals. We modified the standard protocol to target DNA fragments of 330– 400 bp during gel size selection, so that the size of genomic DNA inserts targeted the range 200–270 bp, to produce overlapping paired-end reads for a large proportion of sequenced fragments (Figure 2). We sequenced the RAD libraries in portions of two lanes (grouped with other RAD sequencing experiments) on an Illumina HiSeq sequencer at the University of Oregon, producing 153-bp paired-end reads.

We processed the sequence data and grouped the read pairs from all individuals into RAD loci using several modules from the STACKS software package, version 0.998[37]. First, using the STACKS program process_radtags.pl, we sorted read pairs by barcode, filtered for read quality and removed any pairs in which the forward read did not

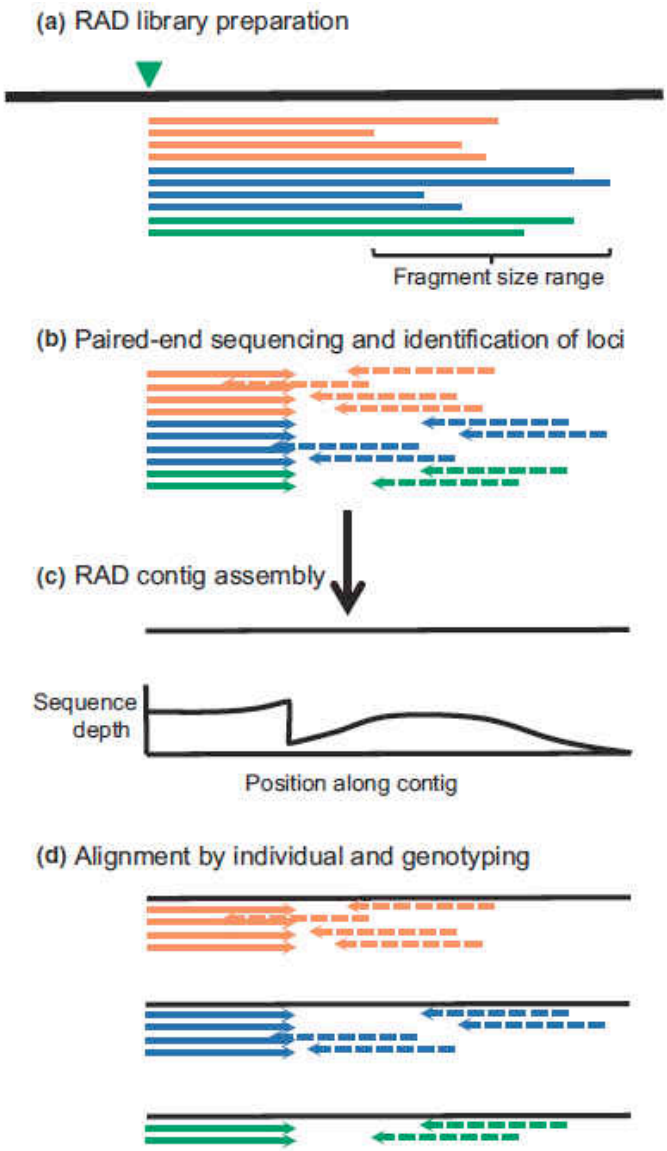

Figure 2. Schematic diagram of overlapping paired-end restriction site-associated DNA (RAD) sequencing. (a) RAD libraries are prepared according to Etter et al. (2011a,b), with the exception that a smaller size range of fragments are selected to obtain overlapping reads. The green triangle indicates the restriction enzyme cut site, and fragments from only one side of the cut site are shown for three individuals (represented by different colours). (b) Libraries are sequenced by Illumina with paired-end reads. Loci are identified with STACKS software, using only the forward reads (solid lines) to cluster reads by locus. (c) Both the forward and reverse reads from each locus are pooled across a set of individuals and assembled into a RAD contig. The depth of sequencing coverage across overlapping paired-end RAD contigs has a unique signature. (d) Reads from each individual are separately aligned against the reference contig set and diploid single nucleotide polymorphism genotypes are called statistically. The length of genotyped sequence data may vary across individuals, and in some cases genotype data may have a gap where paired ends did not overlap.

contain both a correct barcode and the remaining six bases of the SbfI recognition sequence. We then removed read pairs that represented PCR duplicates using the STACKS program clone_filter. The random shearing step in RAD sequencing produces staggered paired-end reads as described previously, so that any set of read pairs that are identical across both the forward and reverse reads are probably PCR duplicates of a single original genomic DNA fragment[31]. Because genotyping depends on using read counts of alternative alleles in a statistical sampling model, PCR duplicates can be misleading because they do not represent independent samples from the genomic pool of DNA.

We identified RAD loci by applying ustacks to the forward reads across all individuals. We enabled the Deleveraging and Removal algorithms to filter out highly repetitive, likely paralogous loci, and we used a maximum nucleotide distance between stacks of 4 to achieve a balance between filtering paralogs and maintaining true alleles at a single locus approximately consistent with the expected number of RAD loci [30,38]. We created a catalog of RAD tag loci using cstacks and matched individuals against the catalog using sstacks. We populated and indexed a MYSQL database of loci using load_ radtags.pl and index_radtags.pl and then exported the data using export_sql.pl. Finally, we grouped the forward and reverse reads from each individual corresponding to each RAD locus using sort_read_pairs.pl.

Contig assembly

We pooled many individuals for contig assembly to increase sequence coverage of read pairs at each RAD locus (Figure 2B). However, we also wanted to limit levels of polymorphism that could complicate assembly. Therefore, we pooled data from 60

individuals from the three populations with the lowest level of admixture as estimated from previous microsatellite data[12]: Lower Hay, Nicola and Tepee. We grouped the forward and reverse reads from all individuals in these populations into a separate file for each RAD locus, using the STACKS program sort_read_pairs.pl. We assembled the reads in each file separately to produce a set of RAD contigs (Figure 2B), using both VELVET[39] and CAP3[40] assembly software. Because CAP3 performed better (see Results), all further analyses mentioned below used the CAP3 assemblies. Because of our pooling strategy, the consensus sequences in this reference set of RAD contigs represent primarily WCT with minimal RBT admixture.

Genotyping and admixture estimates

We aligned the filtered read pairs for each individual from all five populations against the reference set of RAD contigs (Figure 2C). (Three individuals with very low coverage were dropped: one each from Meadow, Nicola and Tepee, leaving a total sample size of 94 individuals.) We used the alignment software BOWTIE[41], allowing up to three nucleotide mismatches in the first 30 bp of each read and up to 15 mismatches over the total read. These parameters represent a compromise aimed at producing valid alignments to the reference, while minimizing bias against divergent RBT haplotypes. We chose them after aligning and genotyping a subset of the data across a wide range of parameter values, but we found that alignment parameters created only marginal differences in overall genotype calls (not shown). We retained only those read pairs that aligned uniquely to the reference contig set and that aligned in the expected orientation (i.e. the forward read aligns at position 0 of the contig, matching the position of the

restriction enzyme cut site and the reverse read aligns in the opposite direction along the same contig within a distance up to 750 bp).

We assigned diploid genotypes to each nucleotide position for each individual using a maximum-likelihood method[42], modified by bounds on the per-nucleotide sequencing error rate of $0.0001 < e < 0.0025$ and a significance level of $a = 0.05$ (custom software available at http://webpages.uidaho. edu/hohenlohe/software.html). These limits have the effect of being more likely to call a heterozygous genotype. While in *de novo* genotyping, these bounds would increase the frequency of false alleles; here, we are genotyping against previously identified WCT and RBT alleles. This strategy and the relatively high significance threshold are also justified because of the quality filtering and removal of PCR duplicates described previously, which increases confidence that each read represents a true independent sample of genomic sequence.

We used previously identified species-diagnostic SNP loci to assess introgression from RBT into these WCT populations. From the RAD sequencing data in WCT and RBT published by Hohenlohe et al.[30], we extracted all RAD loci in which there was either one SNP fixed between species and no other polymorphism in the 54-bp sequence (2923 loci), two fixed SNPs and no other polymorphism (643 loci), or one fixed SNP and one additional SNP polymorphic within either species (1348 loci), for a total of 4914 diagnostic SNPs. We aligned both the WCT and RBT alleles of these 54-bp sequences against the new reference set of RAD contigs, using BOWTIE[41] and allowing up to two nucleotide mismatches. We retained only those diagnostic loci that aligned uniquely with up to two mismatches (for both the RBT and WCT alleles) to the reference contig set.

We then genotyped all individuals from the five admixed WCT populations in the current study as WCT, RBT or heterozygous at each of these loci for which genotype calls were made previously (any genotype calls that did not match previously identified alleles at these SNPs were treated as missing data). As a final filtering step for paralogous loci, we removed loci for which these genotypes exhibited observed heterozygosity >0.5 and FIS < 0.5. Using all such diagnostic SNPs for which at least half of the individuals (47 or more) were genotyped, we estimated proportion of admixture at the locus, individual and population levels as the frequency of RBT alleles across diagnostic loci.

We applied the heterogeneity test of Long[43] to test for super invasive alleles. This analysis tests whether the variance in admixture across loci exceeds that expected from random sampling as well as genetic drift across loci (other tests for admixture outliers do not account for drift and may suffer from a high false positive rate, so our approach is a conservative test[44]). Because this method cannot handle allele frequencies of 0.0, we used Bayesian estimates of allele frequencies with an uninformative prior[44]. We adjusted for differences in sample size of genotypes across loci, which affect the expected variance in allele frequency estimates, in equation 6 of Long[43]. For each locus in each population, we calculated a P-value for the deviation from expected admixture and adjusted for false discovery rate at a level of a = 0.05 within each population[45]. We identified candidate super invasive alleles as those with significantly elevated admixture proportions in two or more populations.

RESULTS

RAD sequencing and contig assembly

After filtering for read quality and presence of a correct barcode and SbfI recognition site, we generated 63,061 577 RAD sequence read pairs across 94 individuals in five admixed WCT populations. Of these, 22% represented PCR duplicates and were removed, leaving 49,248,922 unique read pairs. We identified a total of 222,830 putative RAD loci in STACKS using the forward reads of each pair across all individuals. Only 82,721 of these loci represented eight or more read pairs across all individuals.

We pooled the read pairs corresponding to these 82,721 loci for individuals from three populations with the lowest previously estimated admixture proportions (Lower Hay, Nicola and Tepee). We conducted separate assemblies at each locus using both VELVET[39] and CAP3[40]. In VELVET, we used fixed k-mer lengths of 25, 35, 45 and 55 bp as well as optimizing the k-mer length across these values independently at each locus. All of these assemblies failed to connect overlapping paired-end reads at many loci, and the maximum contig length per locus was only ~100–300 bp. Thus, in many cases, the contigs assembled were smaller than the read length of 147 bp (after trimming the barcode) for the forward reads, meaning that sequences were broken into k-mers and unable to be reassembled. This difficulty in paired-end assembly of RAD data has been observed elsewhere[46], although that study had better success than we did in optimizing assembly parameters per locus. The general problem may be due to the unique signature of sequence coverage expected across contigs for overlapping paired-end RAD data (Figure 2C).

In contrast, the simpler algorithm of CAP3 performed much better. While more computationally intensive, it is still feasible on a desktop computer because the locus identification from STACKS significantly reduces the complexity of each individual assembly. Of the 82,721 loci, 72,124 (87.2%) assembled into single contigs, all but one containing both the overlapping forward and reverse reads. An additional 5,017 loci assembled into two or more contigs, of which only the largest contig was anchored at the expected restriction enzyme recognition site. Of these, all but 151 contained both the forward and reverse reads. We combined these to produce our final reference set of RAD contigs, which contained 77,141 contigs from 82,721 loci (93.3%). Fragment size selection to produce overlapping paired-end reads was remarkably successful, so that over 93% of loci produced contigs spanning the forward and reverse reads. Contig lengths ranged from 147 to 519 bp with most between 250 and 450 bp (Figure 3A), suggesting that longer fragments were carried through the gel-based size selection step. The mean number of read pairs contributing to each contig was 379.3. Contig length was positively related to the number of sequence pairs contributing to each assembly (Figure 3B), so our strategy of pooling individuals to increase coverage at this consensus assembly step appears sound.

Genotyping and admixture

We aligned 54-bp RAD sequences for 4,914 previously identified SNP loci[23,30] against the reference RAD contig set. Of these, 3,456 (70.4%) aligned uniquely to a single contig in the reference set with up to two mismatches for both the RBT and WCT alleles. In addition, 392 (8.0%) aligned to multiple contigs with relatively few

mismatches. These multiple contigs appear to represent genomic regions with duplicate sequence beyond the 54 bp length of the previously identified RAD sequence.

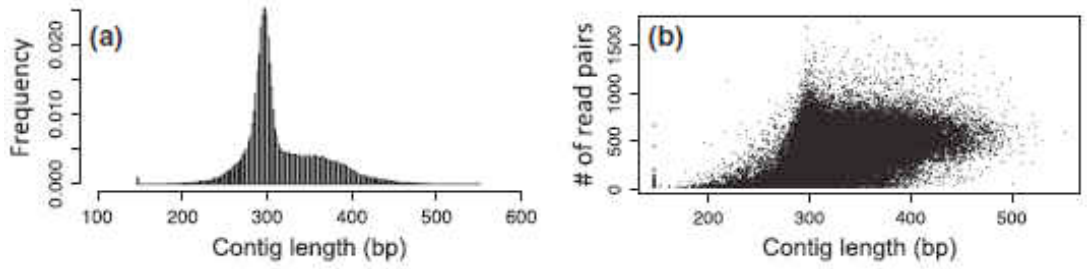

Figure 3. (a) Frequency histogram of consensus sequence lengths across 77 141 contigs assembled by CAP3 from overlapping paired-end restriction-site-associated DNA (RAD) sequencing in admixed westslope cutthroat trout populations. (b) Relationship between sequencing depth at each locus (number of sequence pairs from 60 pooled individuals) and RAD contig length.

We genotyped each individual at all nucleotide positions aligned to the reference contig set using the maximum-likelihood statistical approach described previously. Of the 3456 uniquely aligned diagnostic SNP loci, 3,182 had diploid genotype calls for at least half the individuals sampled. Two of these were probably paralogous loci, with elevated observed heterozygosity (0.95 and 0.80) and reduced FIS (0.90 and 0.61, respectively), and these were removed from further analysis. The remaining 3180 loci had observed heterozygosity <0.45 and FIS > 0.23, suggesting a clear break between them and the two presumptive paralogous loci. We translated genotypes for the final list of 3,180 loci into homozygous WCT, heterozygous or homozygous RBT and assessed proportion of admixture as simply the frequency of RBT alleles.

For all of the individuals genotyped here, we also had individual-level estimates of admixture proportion based on seven species-diagnostic microsatellite loci[12]. Our SNP-based estimates were highly correlated with previous microsatellite-based estimates

overall and within each population, although they tended to be slightly lower (Figure 4A).

We detected evidence of introgression in all 53 individuals for which no RBT alleles had been observed at the microsatellite loci. In these individuals, RBT alleles were detected at 1–235 loci, leading to individual admixture proportions ranging from 0.0013 to 0.0439 that were undetected in the microsatellite data. Average population- level admixture proportions are also consistent with microsatellite-based estimates (Pearson r = 0.99; P = 0.0013; Figure 4B), in which Dutch and Meadow exhibited higher levels of admixture than the other three populations, although SNP-based estimates were lower than microsatellite estimates for four of the five populations.

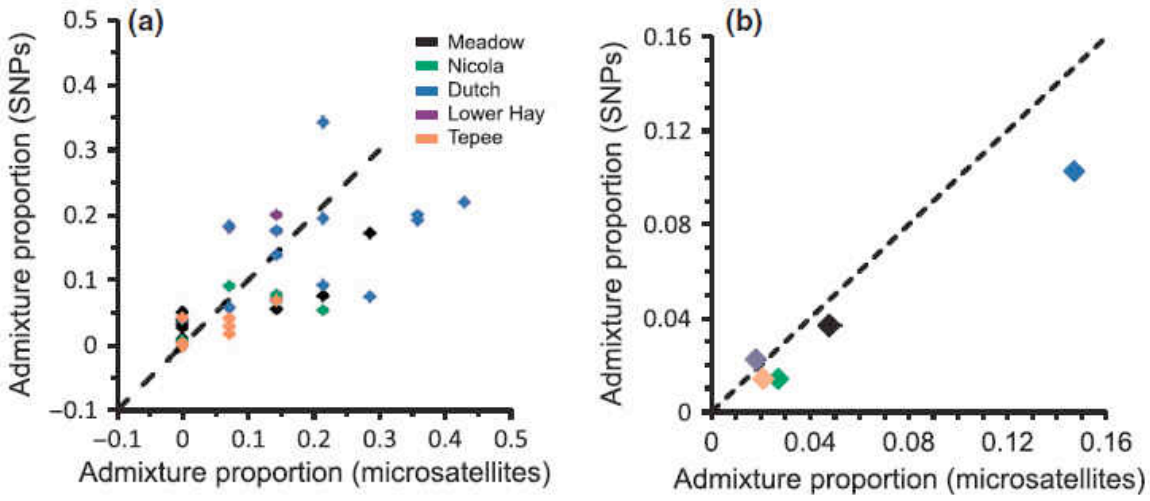

Figure 4. (a) Individual-level admixture proportions estimated from seven diagnostic microsatellite loci (Boyer et al. 2008) vs. current estimates from 3180 single nucleotide polymorphism loci across 94 westslope cutthroat trout individuals from five populations. Note that many of the points, particularly those with admixture proportions near 0.0, lie on top of each other. (b) Population-level admixture proportions estimated from the same two data sets, calculated using only the individuals genotyped by both Boyer et al. (2008) and the current study.

Comparing admixture proportions across SNP loci reveals a positively skewed distribution within each population and overall, with many loci showing little or no

admixture and a small set of outlier loci (Figure 5). Of the 3,180 diagnostic SNP loci genotyped, 634 showed no RBT alleles in any of the five populations. However, 94 loci exhibited admixture levels of 0.1 or greater across all five populations combined, up to a maximum of 0.542 (Figure 5F). These are candidate super invasive alleles: RBT alleles that may have spread rapidly or have higher probabilities of persistence in WCT populations. Within each population, loci exhibited significantly elevated admixture proportions using the heterogeneity test of Long[42], corrected for false discovery rate (Table 1). Three loci were significantly invasive in two or more populations, one of which was significant across all five populations (Figure 5).

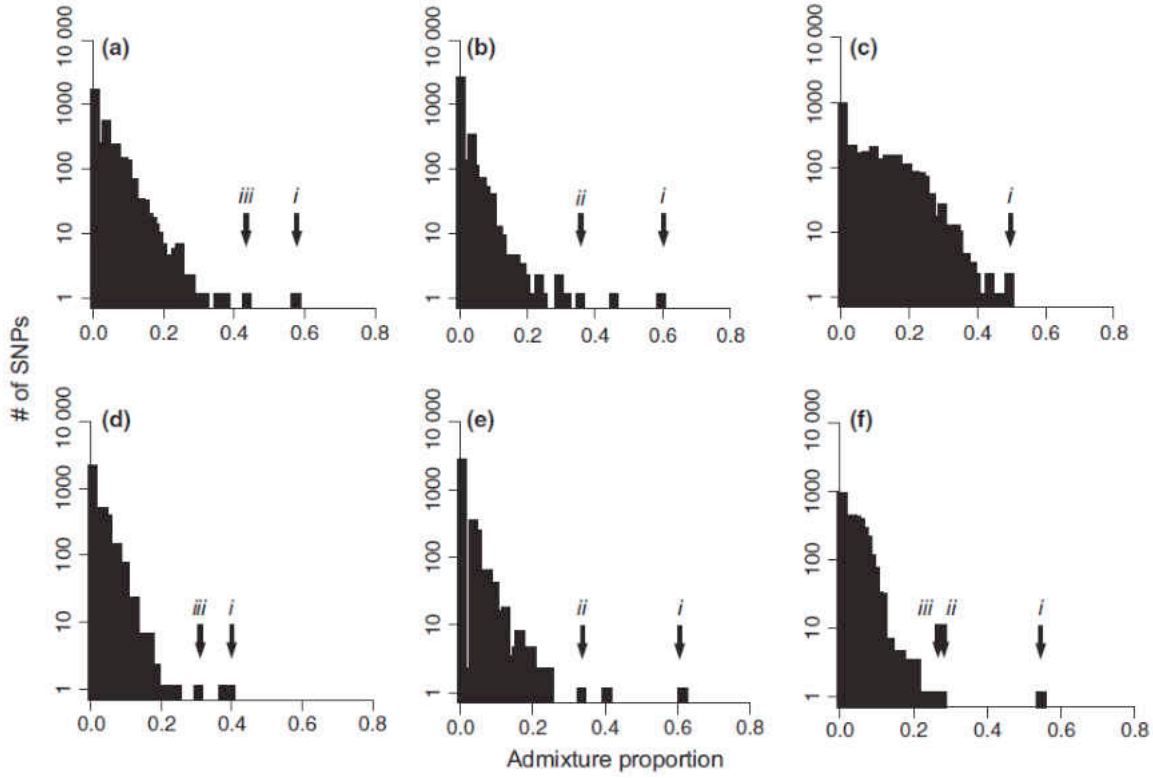

Figure 5. Frequency histograms of admixture proportion across 3180 diagnostic single nucleotide polymorphism loci. (a) Meadow. (b)Nicola. (c) Dutch. (d) Lower Hay. (e) Tepee. (f) All five populations combined. Arrows indicate super invasive alleles—loci with significantly elevated admixture proportion (a = 0.05, corrected for false discovery rate) independently in two or more populations.(i) restriction-site-associated DNA (RAD) locus 118904. (ii) RAD locus 117399. (iii) RAD locus 82847.

Table 1. Correlation between previous microsatellite and current single nucleotide polymorphism (SNP)-based estimates of individual-level admixture proportions, and super invasive alleles exhibiting significantly elevated introgression with a false discovery rate corrected *P*-value.

Genotype confidence threshold	Total number of SNP	Error A		Error B	
		No. parental genotype combinations 00/00 or 11/11	%offspring with error	No. parental genotype combinations 00/11 or 11/00	%offspring with error
0	6431	3558	12.4	65	18.5
1	4748	3071	7.2	50	6
2	3779	2716	3.2	49	4.1
3	2930	2270	1.5	40	0
4	2579	2055	0.9	38	0
5	2115	1697	0.5	34	0
6	1898	1534	0.3	32	0
7	1559	1253	0.2	27	0
8	1409	1131	0	26	0

We conducted a translated nucleotide BLAST search using the RAD contig sequence for each of these three super invasive alleles. Two of them aligned closely to annotated genes whose function is consistent with selection in hybridized WCT populations. The locus significantly admixed in all five populations (RAD locus 118,904) aligned significantly to the vertebrate gene latent transforming growth factor beta-binding protein 2 (LTBP2), with the most significant hit in Bos taurus (E-value = 10 e-7). The second locus, significantly admixed in the Nicola and Tepee populations (RAD locus 117,399), aligned to the vertebrate gene furry homolog-like (FRYL), with the most significant hit in zebrafish (Danio rerio, E-value = 10 e-9). It is worth noting that the BLAST alignments to these two annotated gene sequences began at nucleotide positions 191 and 210, respectively, of the RAD contigs so that the identification of these candidate genes would not have been possible solely with single-end RAD sequence data.

DISCUSSION

Genomic tools hold remarkable promise for conservation and management of many taxa. The ability to rapidly identify and genotype large numbers of genetic markers allows improved estimates of demographic parameters (gene flow, effective population size, population-level admixture), as well as identification of outlier loci (locally adapted genes, invasive alleles). Overlapping paired-end RAD sequencing offers advantages for rapid development of large numbers of candidate SNPs that can be used in high-throughput genotyping assays, particularly in the case of large or repetitive genomes.

In a specific application of this technique, here we assessed genomic patterns of introgression and were able to detect individuals with very low levels of admixture, precisely estimate individual- and population-level admixture and detect candidate super invasive alleles driven to high frequency by selection. Below, we discuss some general aspects of the sequencing technique for conservation genomics and lessons from its application to the genomics of hybridization.

Overlapping paired-end RAD for conservation genomics

By assembling contigs of 400 bp or more adjacent to RAD loci, overlapping paired-end RAD provides sufficient flanking sequence for SNP assay design simultaneous with SNP discovery. The ability to generate sufficient flanking sequence has previously been a limitation of RAD sequencing for converting rapid SNP discovery to a set of high-throughput assays[23,47]. Our approach can rapidly provide a multitude of candidate SNP markers for high-throughput assay development. Here, we only analysed a few thousand diagnostic markers that had been previously identified. In general, the

majority of contigs of 300–400 bp or longer would be expected to contain SNPs relevant for most population genomic or conservation applications.

Assembling RAD contigs provides more continuous genomic sequence data for discriminating paralogous loci. This is a particular challenge in salmonids because of their ancestral genome duplication, which created homeologous duplicate sequence across the genome[26, 27, 48]. Here we found examples of loci sharing very similar sequence over ~50 bp, so that they were grouped together in previous analysis, but diverged beyond that length. As a result, we were able to further screen the candidate diagnostic SNP loci we had previously identified[23,30] by removing the 8% that aligned to multiple RAD contigs. Ongoing validation of the reduced set will determine the success rate of these refined candidate markers.

Our approach to RAD contig assembly produced a single contig with high average read depth for most of our RAD loci. Nonetheless, the assembly and validation of RAD contigs can be challenging[46]. Assemblies using the de Bruijn graph technique of VELVET[39] produced consistently shorter contigs than a simpler (but more computationally intensive) assembly algorithm in CAP3[40]. This contrasts with the results of Etter et al.[33,36], who had better success with VELVET in assembling the reverse reads from nonoverlapping paired-end RAD. Willing et al.[34] used nonoverlapping paired-end RAD in guppies and assembled the reverse reads for 91.3% of loci into a single contig with generally lower sequence coverage than used at the assembly step here. That study used the assembler LOCAS, specifically designed by one of the authors for low-coverage data. Davey et al.[46] had poor results with LOCASOPT and VELVET in assembling paired-end RAD data from Heliconius butterflies, but better results using the

computationally intensive VELVETOPTIMISER. In our trout data set, over 87% of loci produced a single contig of both forward and reverse reads with CAP3, and many of the remainder could be filtered out as paralogs.

Techniques like overlapping paired-end RAD sequencing may allow new analytical power. Compared with other markers like microsatellites, SNPs can be limiting in that they typically exhibit only two alleles in natural populations. More power to understand population genetic processes would come from using multi-allelic haplotypes instead of SNPs in analyses of high-throughput sequence data[49,50]. Because of the relatively long contigs that can be generated[33, 34, 36] and because haplotype phase is known across read pairs and thus can be inferred along the length of RAD contigs, paired-end RAD offers the possibility of using haplotype- rather than SNP-based analyses. Genealogical relationships among multiple haplotypes are very useful for inferring demographic and evolutionary history[51,52].

Assessing genome-wide patterns of introgression

Here we provide one of the first genome-wide assessments of human-mediated introgressive hybridization in salmonid fishes (see also Lamaze et al.[53]). Our results confirm previous patterns of hybridization between introduced RBT and native WCT in the North Fork Flathead system[12, 13]. Population-level admixture estimates were generally consistent for diagnostic microsatellites and RAD-based SNP loci, suggesting that thousands of diagnostic loci are generally unnecessary for approximate estimates of population- level admixture. However, one estimate did differ: the estimate for Dutch Creek was over 40% higher using the microsatellite data (Figure 4B). This may be explained by selection against RBT alleles in chromosomal regions near RAD loci and/or

sampling error from using only seven diagnostic microsatellite loci, especially for populations with low levels of introgression. Given the variation in introgression we observed here among SNP loci, the genomic location of those microsatellite loci could also be a major source of variation.

Overestimation of admixture (by using only a handful of neutral loci) could cause populations to not be protected under conservation laws, such as the U.S. Endangered Species Act (ESA). For Lahontan cutthroat trout, listed under the ESA, 10% RBT admixture is the threshold for a population to be protected as if it were nonhybridized (pure native) Lahontan. Based on sampling theory for neutral loci, it is likely that 50–100 diagnostic loci would improve accuracy to levels approaching that of thousands of RAD loci, if those diagnostic loci are widely distributed across the genome[23].

At the individual level, overlapping paired-end RAD sequencing allowed detection of very low levels of RBT introgression. Here, we detected RBT alleles in all 94 samples analyzed, over half of which did not exhibit RBT alleles at seven microsatellite loci[12]. Some of the assumed RBT-diagnostic alleles could actually exist in non-hybridized WCT populations. Additional RAD sequencing of pure-native populations (e.g. isolated above barriers in the Flathead River) could help identify assumed diagnostic RBT alleles that might exist in WCT (e.g. due to maintenance of ancestral polymorphism).

Genome-wide marker coverage is an important advance for conservation and management because it allows powerful screening of individuals to prevent inadvertent release of hybridized individuals into populations (e.g. during assisted migration, brood stock development, translocation and reintroduction) and identification of markers for

rapid screening for early detection of hybridization. From a landscape genetics perspective, the ability to precisely estimate admixture would allow fine spatial mapping of hybridization and introgression patterns. This approach may be useful in monitoring and preventing the spread of invasive species and their alleles in many plant and animal species facing hybridization threats in nature[54].

Dense coverage of markers across the genome allows for detection of candidate super invasive alleles—alleles of an invasive taxon that rise to much higher frequency (level of introgression) than the genomic background, analogous to outlier loci in genome scans for selection[55]. Here we detected several candidate super invasive alleles as evidenced by the distributions of admixture proportions among SNPs (in all populations) containing a long tail of outlier loci. Several of these loci were consistent as outliers across populations. Further study is needed to confirm that these are indeed RBT alleles that have introgressed into these WCT populations. The haplotype information provided by longer overlapping paired-end RAD (e.g. using 250-bp reads as provided by Illumina MiSeq technology) may facilitate that analysis. Further study would also be needed to identify the phenotypic and fitness consequences of these invasive alleles.

BLAST searching revealed close sequence matches for two candidate invasive alleles to vertebrate genes (LTBP-2 and FRYL). Super invasive alleles may be under positive selection and increase fitness in hybridized populations. Alternatively, they may spread by having phenotypic effects on dispersal or through segregation distortion, despite reducing overall fitness from outbreeding depression[56]. The LTBP family of proteins interacts with TGF-beta and has a wide range of developmental and physiological functions, including effects on fertility[57-59], although the specific

relationship between LTBP-2 and TGF-beta is unclear[60]. In RBT, the related protein LTBP-3 and other related proteins have been implicated in early ovarian development and early embryonic development[61-63], suggesting the hypothesis that the RBT allele at this locus positively affects fecundity in admixed individuals. It is exciting that future research and additional studies like this one will help understand mechanisms driving super invasive alleles and genome-wide introgression in natural populations.

BRIDGE TO CHAPTER IV

We have seen in chapters II and III that restriction site flanking DNA can be used to probe genomic sequence. Chapter IV describes work that inverts this process by using known genomic sequence to assay restriction enzyme specificity. Modifications to the molecular RAD site preparation allow quantification of both cleavage sites and flanking preferences in type II restriction endonucleases. This is done using sequenced genomes as substrates in order to map sequencing events back to their cleavage sites in a countable fashion. Sequenced genomes provide easily acquired complex substrates that contain all possible short nucleotide stretches in known number and context. The methods described in chapter IV leverage these previously mapped templates to quantify restriction enzyme activity in a massively parallel fashion.

CHAPTER IV

MASSIVELY PARALLEL RESTRICTION ENZYME ASSAY

This work was published in volume 189 of the journal of *Nucleic Acids Research* in 2013 in collaboaration with Quimby A, Zhu Z, and Johnson EA. I was responsible for experimental design, bench work, analysis, and manuscript preparation.

INTRODUCTION

Type II restriction endonucleases cleave double stranded DNA at a constant position with respect to a short (3-8bp) recognition sequence[1]. Their exquisite specificity has rendered them among the most useful tools in molecular biology[1,2]. However, the impact of additional variables such as organic solvent, ion, small molecule and enzyme concentrations has large effects on the specificity of restriction endonucleases, often leading to cleavage at non-cognate sites (termed star activity)[3-7]. Many commonly used restriction endonucleases show some star activity even under standard reaction conditions[3]. The DNA substrate itself can also modulate cleavage. It has been noted that nucleotides flanking the recognition site confer large contributions to the energetics of cleavage[8-12]. Quantitative analysis of star activity and flanking effects will help to elucidate the structure-function rules for restriction enzymes, define the window of optimal restriction endonuclease specificity as well as tailor reaction conditions toward novel target sequences.

Despite the conserved functionality amongst the restriction endonuclease family, these enzymes show great divergence in both sequence and mechanism[1,9,13,14]. Apart from isoschizomers, most members show little sequence homology to each other or other

known proteins[1]. Additionally, the variable distribution of base-contacting residues amongst the restriction endonucleases has confounded recognition sequence prediction[9,15,16]. Consequently restriction endonuclease characterization must be carried out empirically for each enzyme. Star activity[4,17-21] and flanking preference[8-12] have been investigated for several enzymes. These experiments have been performed on homogeneous substrates. A series of oligonucleotides containing different star or flanking sequences are synthesized, annealed, cleaved and analyzed one by one, making exhaustive studies difficult. Recognition site determination is typically carried out by digestion of a homogeneous plasmid or virus DNA substrate followed by agarose gel visualization of cleavage products[6,22-25]. This technique is lacking both in its substrate complexity and sensitivity. A given cognate or star site could occur very few times in these substrates, and at times not at all. This limits the ability to accurately quantify activity at different cleavage sites due to a lack of diversity of flanking nucleotides. Star activity is often several orders of magnitude lower than cleavage at the cognate site[3,17]. Consequently a large component of star activity will remain cryptic when cleavage products must be of sufficient abundance to be visualized on an agarose gel.

The growing amount of prokaryotic genomic sequence putatively coding for uncharacterized restriction endonucleases[26,27] in conjunction with ongoing efforts to engineer altered specificities[22-25,28-30] will be aided by high-throughput methods to quantify restriction endonuclease activity instead of the methods currently available. For example, in order to characterize the genome-wide digestion patterns of the methylation-specific restriction endonuclease AbaSDFI[31], genomic rat brain DNA was digested with AbaSDFI to map 5-hydroxymethylcytosines, and the digestion products were cloned into

plasmids and Sanger sequenced one by one to map 122 cleavage sites to the rat genome. A similar strategy was used to demonstrate the relaxed specificity of the restriction enzyme TspGWI in the presence of sinefungin by Sanger sequencing 218 clones[5].

High-throughput sequencing has become a valuable tool for analyzing DNA-protein interactions. The ability to experimentally pair a DNA-protein interaction to a sequencing event has enabled techniques such as ChIP-seq[32] to provide sensitive statistics on transcription factor-DNA binding. We use derivations of the RAD-seq[33] method to quantitatively measure restriction endonuclease activity across the sequenced *D. melanogaster* and *E. coli* genomes. This method specifically prepares DNA adjacent to restriction sites for Illumina sequencing, allowing the relative sequence counts of sites with different flanking nucleotides to be determined. The RAD-Seq protocol was carried out with serial enzyme dilutions to identify flanking motif enrichment in enzyme-limiting reactions. Modifications were made to the protocol to sequence all cleavage events regardless of overhang in order to generate a complex profile of relative activities at cognate and star sites in a single experiment. We apply these methods to quantify the cleavage patterns of EcoRI and MfeI and to compare star activity with their engineered high-fidelity counterparts, and to quantify the effect of flanking nucleotides on MfeI activity.

MATERIAL AND METHODS

All enzymes and buffers were contributed by New England BioLabs.

Star activity assay

To assay restriction enzyme activity on a genome-wide scale we designed an unbiased strategy to sequence all digested fragments regardless of overhang (Figure 1). 1000-1500 base pair fragments of *E. coli* strain REL606 DNA were digested under star conditions, and smaller 300-500 base pair fragments whose decreased molecular mass indicated digestion were separated and Illumina sequenced. 1,000,000 reads from both EcoRI digests and 650,000 reads from both MfeI digests were mapped back to the REL606 genome, and adjacent cleavage sites were computationally analyzed.

(1) Generation of random 1000-1500 base pair digestion templates: 3 ug of REL606 genomic DNA were randomly sheared by sonication (Bioruptor). DNA fragments between 1000-1500 base pairs were then separated and purified by agarose gel electrophoresis. The DNA was then blunt end repaired using Quick Blunting Kit and 3' adenylated using Klenow exo^{-}. To distinguish the sheared DNA ends, non-divergent Illumina end 2 adapters composed of annealed oligonucleotides 5'-Phos-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG-3' and 5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT-3' were ligated to the 1000-1500 base pair pool using concentrated T4 DNA ligase. 10 ng of this sample were used in a 20 cycle Phusion polymerase PCR reaction with the Illumina PE primer 2.0 (5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT-3') following Phusion product guidelines to select 1000-1500 base pair fragments with the Illumina end 2 sequence on each end.

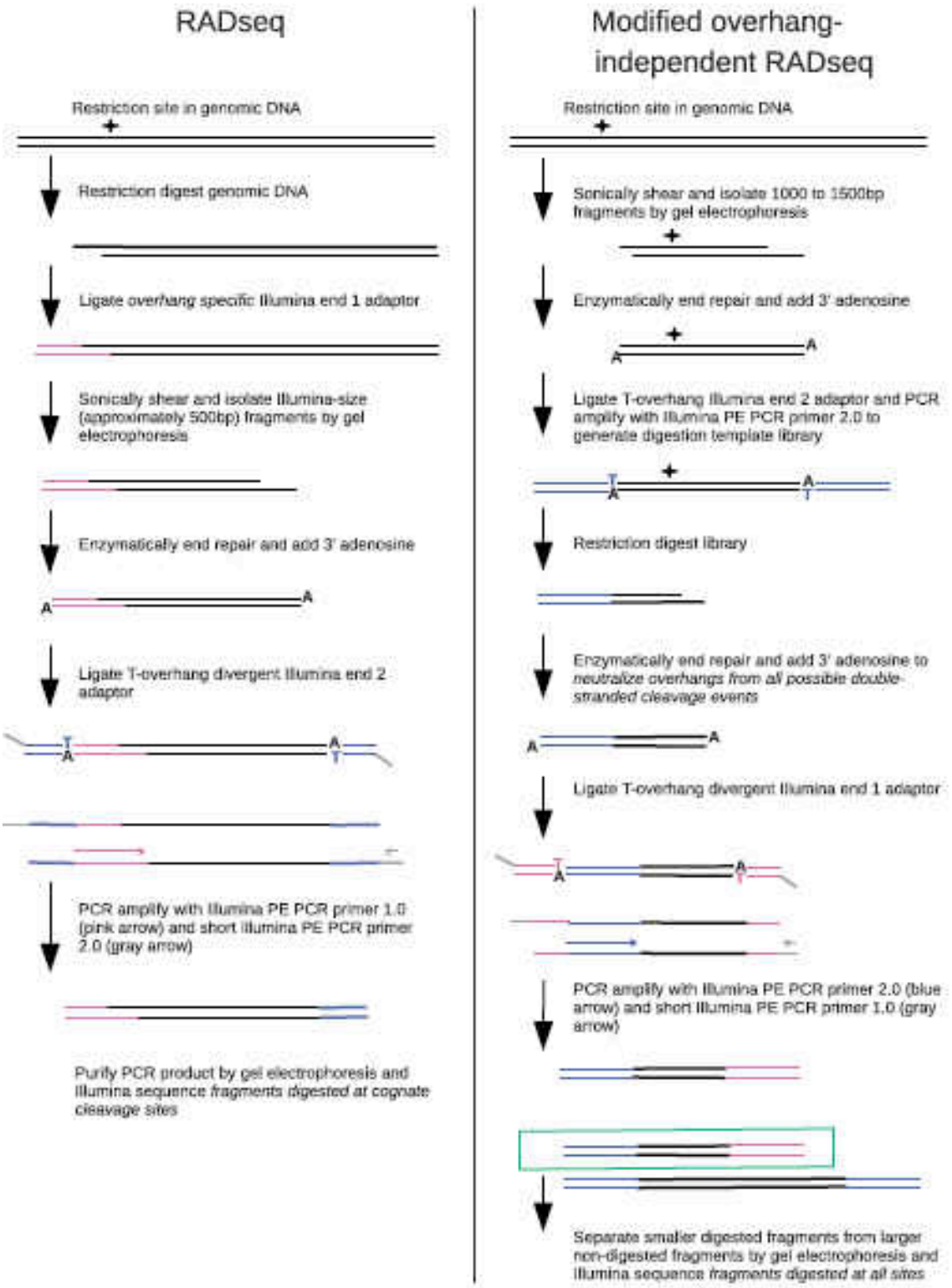

Figure 1. The path of a single restriction site–containing genomic locus is shown for both the RAD-seq protocol (left) and the modified overhang-independent RAD-seq protocol used in the star activity assay (right).

(2) Star condition digest: In order to generate a complex cleavage activity profile, DNA from the previous step was digested with an excess of restriction enzyme. 52 ng of DNA was digested with 50 units of MfeI (GenBank accession number SRR652142) or MfeI-HF (accession SRR652141) in a 50 uL reaction containing 1X NEB4 and 5% glycerol for 24 hours at 37°C. 32 ng of DNA was digested with 200 units of EcoRI (accession SRR652140) in a 50 uL reaction containing 1X NEB1 and 10% glycerol for 24 hours at 37°C. 32 ng of DNA was digested with 200 units of EcoRI-HF (accession SRR652139) in a 50 uL reaction containing 1X NEB4 and 10% glycerol for 24 hours at 37°C.

(3) Tagging of cleaved end with Illumina end 1 adapter: The digested DNA was blunt end repaired using Quick Blunting Kit to neutralize all potential overhangs. The DNA was then 3' adenylated using Klenow exo$^-$ and ligated using concentrated T4 DNA Ligase to barcoded divergent first-end Illumina adapters composed of annealed oligos 5'-GGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-barcode-T-3' and 5'-Phos-barcode-AGATCGGAAGAGCGTCGTGTACTACGTT-3'. 10 ng of DNA from the ligation reaction was then used as template for an 18 cycle Phusion PCR reaction with Illumina primers PE PCR Primer 1.0 (5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3') and PE PCR Primer 2.0 (5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT-3') following Phusion product guidelines. Use of a divergent first-end adapter requires the paired-end primer to anneal first for amplification to occur. This eliminates the first end sequence on the sheared side. We used change in molecular mass to select for digested molecules. As the pre-digestion sample ranged from 1000-1500

base pairs, we agarose gel-purified 300-500 base pair PCR fragments for sequencing to assure cleavage.

This final library exclusively contained molecules with a first end Illumina sequence on the cleaved side and a second end sequence on the sheared side. The experimental samples were sequenced on an Illumina HiSeq 2000 to generate 100 base pair single end reads beginning at the cleavage site. The samples were separated by barcode and the reads were mapped back to the *E. coli* genome to infer the cleavage site.

Fidelity index determination

The fidelity index (FI) was determined for EcoRI-HF and MfeI-HF by the standard method[3]. The substrate used for all FI determinations was lambda DNA.

Flanking sequence preference assay

To determine the flanking sequence preferences of MfeI, *Drosophila melanogaster* genomic DNA was digested in saturating and enzyme-limiting conditions. The DNA adjacent to the restriction sites was then PCR amplified and Illumina sequenced as per the RAD-Seq protocol[33]. 550,000 reads from each digest were mapped to the *Drosophila* genome. Flanking sequence preference was inferred from motif enrichment in enzyme-limiting conditions.

1) MfeI digests: Digests were carried out in 50 uL reactions containing 786 ng of *D. melanogaster* strain Oregon-R genomic DNA, 1x NEB 4, 1% glycerol and varying amounts of MfeI for 15 minutes at 37°C. A range of partial digest conditions was achieved by varying the amount of enzyme through 12 serial dilutions each decreasing enzyme concentration by a factor of two as follows: Reaction 1 contained 10 units (GenBank accession number SRR652186), Rxn 2: 5 units (accession SRR652187), Rxn

3: 2.5 units (accession SRR652188), Rxn 4 : 1.25 units (accession SRR652189), Rxn 5: 0.63 units (accession SRR652190), Rxn 6: 0.31 units (accession SRR652191), Rxn 7: 0.16 units (accession SRR652192), Rxn 8: 0.08 units (accession SRR652193), Rxn 9: 0.04 units (accession SRR652194), Rxn 10: 0.02 units (accession SRR652195), Rxn 11: 0.01 units (accession SRR652196), Rxn 12: 0.005 units (accession SRR652197).

2) RAD-Seq library preparation: RAD-Seq libraries were prepared according to Baird *et al.*[33] with the following parameters. MfeI RAD adapters were composed of annealed oligonucleotides of the form 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCC TACACGACGCTCTTCCGATCT-barcode-3′ and 5′-Phos-AATT-barcode-AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGT ATCATT-3′. Each of the 12 experimental digests was ligated to an MfeI RAD adapter with a unique barcode to allow sequencing on the same Illumina HiSeq 2000 lane. Prior to amplification, reactions 1-4 (high-enzyme), 5-8 (mid-enzyme), and 9-12 (low-enzyme) were pooled to increase the sequence contribution of the lower enzyme samples as the concentration of digested fragments was expected to be much greater in the higher enzyme samples. The final step in the RAD-Seq library preparation protocol is a PCR enrichment of DNA restriction fragments flanked by both sequences necessary for Illumina sequencing. 10 ng of the high-enzyme ligation were PCR amplified using Phusion polymerase with PE PCR Primer 1.0 and a shortened PE PCR Primer 2.0 (5'-CAAGCAGAAGACGGCATACGA-3') for 15 cycles. This was increased to 17 cycles with 15 ng ligation template for mid-enzyme libraries and 20 cycles with 20 ng template for low-enzyme libraries.

Fragments averaging 550 base pairs were agarose gel-purified from each reaction and sequenced on an Illumina HiSeq 2000 to generate single end 100 base pair reads.

Data processing

Sequence reads were aligned to *Drosophila melanogaster* genome build 5.4.52, or *E. coli* genome REL606 using Novoalign v2.07 (Novocraft.com). Custom Perl scripts counted the sequence reads at each genomic location. For the flanking nucleotide assay, the flanking nucleotides were inferred from the genome reference sequence for each aligned read, and the total counts of reads for each flanking sequence tracked. For the star activity assay, the reads found for each recognition sequence was normalized by their count in the genome.

RESULTS

Star activity assay

Restriction enzymes are known to digest DNA at non-cognate sequences called star sites. We developed a star activity assay for quantifying the relative activity of restriction enzymes at cognate and non-cognate sites using genomic DNA as a substrate. The star activity assay comprises shearing genomic DNA to a defined length, digestion with a restriction enzyme, and selecting amplified fragments much smaller than the original sheared fragments for sequencing. Because the DNA fragments are blunted after digestion, the sequencing adapters ligate equally well to cognate and non-cognate sites. The full sequence of the digested site can be recovered after alignment of the sequence read back to the reference genome. Thus, the relative sequencing coverage of each genomic locus can be quantified, and the normalized sequencing coverage of each

particular site sequence motif, represented many times across a genome, can be determined.

MfeI star activity

After digestion of *E. coli* genomic DNA with MfeI in star activity conditions for 24 hours, non-cognate sequences with single base pair changes from the cognate CAATTG were seen at digested sites. The bulk of non-cognate reads came from CAACTG and its reverse complement CAGTTG, and a small number of additional reads were created by digestion of CAATTA, CAATTC, CACTTG, and their reverse complements TAATTG, GAATTG, and CAAGTG (see Table 1). These star sites were also seen after digestion with an engineered high-specificity version of MfeI (MfeI-HF, NEB), although at much lower coverage compared to wild type MfeI (see Table 1). For example, the percent of total reads for the most abundant star site, CAACTG, was more than six-fold higher for MfeI compared to MfeI-HF. MfeI-HF also showed a substantial reduction in star activity compared to the wild type enzyme when fidelity index was used as the metric. The FI determined in this study of MfeI-HF in NEB4 is over 500, while the previously determined FI of wild type MfeI in NEB4 is only 32[3], demonstrating a greater than 16-fold reduction in star activity on a simple substrate. While both of these assays demonstrate the increased fidelity of the engineered MfeI-HF, their results cannot be quantitatively compared due to differences in substrate and reaction conditions.

EcoRI star activity

After digestion of *E. coli* genomic DNA with EcoRI in star activity conditions for 24 hours, six non-cognate sequences with single base pair changes from the cognate GAATTC were seen at digested sites, although three of these made up a very small

Table 1. Percent of reads at star sites after digestion with MfeI.

Site	MfeI	MfeI-HF
CAACTG	3.03	0.46
CAATTA	0.14	0.07
CAATTC	0.08	0.07
CAAGTG	0.04	0.11

fraction of the reads (see Table 2). These star sites comprised a significant portion of all sequences from the EcoRI digestion, with more than 31% of all reads coming from GAATTT sites, and GAAGTC and GAATTA sites having 4% and 2% of all reads, respectively. The sites GAACTC, GAATTG, and GAATGC together made up only ~0.4% of all reads. The high fidelity version of EcoRI had much improved specificity in star activity conditions. The percent of total reads coming from star sites was 3,000-fold lower in EcoRI-HF compared to EcoRI (see Table 2). As in the comparison of MfeI to MfeI-HF, the coverage difference between EcoRI and EcoRI-HF was less pronounced with the minor-frequency star sites. FI testing also showed a drastic improvement in specificity for the engineered EcoRI-HF. The FI determined in this study of EcoRI-HF in NEB4 is over 16000, while the previously determined FI of wild type EcoRI in NEB4 is only 4^3, demonstrating a greater than 4000-fold reduction in star activity on lambda DNA.

Flanking sequence preference of MfeI at cognate site CAATTG

We also examined how the digestion of cognate sites is affected by the flanking nucleotide sequence. The simpler restriction-site associated DNA (RAD) method was used to generate short DNA tags at each cognate cleavage site. As in the previous assay, the number of tags found at each locus was used as a measure of digestion efficiency. By

Table 2. Percent of reads at star sites after digestion with EcoRI.

Site	EcoRI	EcoRI-HF
GAATTT	31.58	0.01
GAAGTC	4.17	0.01
GAATTA	2.64	0.01
GAACTC	0.31	0.01
GAATTG	0.05	0.01
GAATGC	0.04	0.01

calculating a normalized coverage for each particular flanking sequence the influence of these sequences on restriction enzyme activity could be determined.

We digested genomic DNA from *Drosophila melanogaster* with the restriction enzyme MfeI using enzyme concentrations that ranged from fully saturating to very limiting (10 – 0.005 units). We reasoned that the highest enzyme concentrations would digest every available cognate site to near completion, whereas enzyme preferences for particular sequences in the flanking nucleotides would be apparent at the lowest concentrations. RAD libraries were made for each enzyme concentration and sequenced to an average of ~3X coverage for all sites. The sequence reads were mapped to the genomic sequence and the flanking nucleotides extracted for each site. The read counts for sites or half-sites sharing a flanking sequence were binned, and the average coverage calculated.

The single nucleotide adjacent to MfeI had a strong effect on site preference (see Figure 2A). As the amount of MfeI was diluted, the sequencing reads became concentrated on preferred sites, creating higher coverage depth for preferred sites and lower coverage depth for sites that were digested less efficiently. If the site sequences are

ranked by the change in sequencing coverage from the most enzyme to the least, the greatest increase in coverage is the palindromic GCAATTGC, and the greatest decrease is the palindromic ACAATTGT. In general, there is very strong concordance in the coverage change for sites that are reverse-complements of each other, as would be expected (see Table 3). All the sites with a 5' G base or 3' C base have an increase in coverage under dilute conditions, demonstrating that MfeI has a strong preference for these nucleotides adjacent to the cognate cut site. A 5' T base or 3' A base has a near neutral effect on coverage, and 5' A or C bases or 3' T or G bases have a negative effect on coverage, demonstrating that their presence in the flanking sequence makes an MfeI restriction site less likely to be cleaved in dilute enzyme conditions.

This preference for certain sequences by the MfeI restriction enzyme extends beyond the single adjacent base. The 5' G base preference becomes even more pronounced when the dinucleotide is 5' (A/T)G, but the 5' (G/C)G dinucleotide has a reduced sequencing coverage (see Figure 2B). The preference for A or T bases in the second 5' position away from the cut site is also true for the (A/T)T versus (G/C)T dinucleotides (see Figure 2C), but dinucleotides with a 5' A or C base adjacent to the restriction cut site have more complicated interactions. The TA dinucleotide has a very strong positive effect on sequencing coverage, while (A/C/G)A are all weakly to strongly negative (see Figure 2C). The TC dinucleotide has the lowest sequencing coverage of all dinucleotides, while (A/C/G)C have only a weak effect on coverage (see Figure 2C). Our data show the flanking effects of each dinucleotide to operate independently of the sum of its parts.

Table 3. The change in sequencing coverage from enzyme saturating to limiting conditions for each of the 16 single-nucleotide flanking pairs surrounding the cognate. MfeI site

Site	Change in coverage
GCAATTGC	2.6
TCAATTGC	1.3
GCAATTGA	1.1
GCAATTGG	0.5
CCAATTGC	0.4
ACAATTGC	0.2
GCAATTGT	0.1
TCAATTGA	−0.1
CCAATTGA	−0.8
ACAATTGA	−0.8
TCAATTGT	−0.9
TCAATTGG	−0.9
CCAATTGT	−1.0
ACAATTGG	−1.2
CCAATTGG	−1.4
ACAATTGT	−1.4

Flanking sequence preference of MfeI at star site CAACTG

The abundant non-cognate site CAACTG identified in wild type MfeI star activity conditions was analyzed for flanking nucleotide preferences in order to compare flanking effects in star versus cognate activity. There was a wide range of site sequencing coverage depending on the flanking nucleotide sequence of the CAACTG site. There was a 2.3-fold difference in coverage between sites with a 5' T base and the poorly cut star sites with a 5' A base (see Figure 3) which is of larger magnitude than the single base flanking effects seen at the cognate site. Interestingly, the effect of particular flanking sequences differed between the cognate and star sites. The 5' G base was most preferred

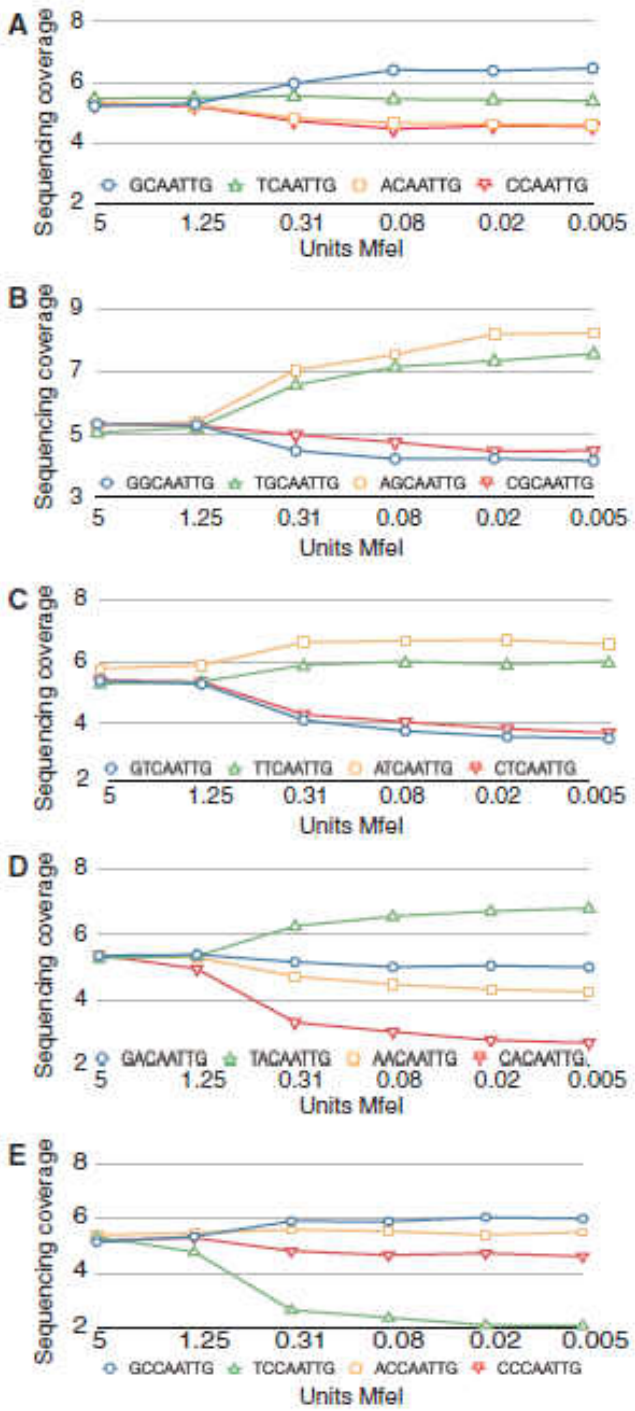

Figure 2. MfeI activity is affected by flanking base preference. All graphs plot normalized sequencing coverage (y-axis) versus units of enzyme (x-axis). Blue circles, G base; green triangles, T base; yellow squares, A base; red triangles, C base. (A) Changes in sequencing coverage for the different bases adjacent to the MfeI half site, i.e. NCAA. (B–E) Changes in sequencing coverage for the different distal bases of the dinucleotide adjacent to the MfeI half site, for the dinucleotide NG-CAA (graph B), NT-CAA (graph C), NA-CAA (graph D), NC-CAA (graph E).

by the cognate site, whereas a 5' T base was most preferred by the star site. The effect of flanking sequences also differed for the two distinct half sites of the CAACTG star site. Whereas palindromic 5' and 3' flanking sequences about the cognate MfeI sites confer the same effect, the distinct star half sites CAA and CAG respond differently. While both preferred a 5' T base and a 5' A base reduced sequencing coverage the most, the next most preferred 5' base was a C base for the CAA half site and a G base for the CAG half site (see Figure 3). Our data show that MfeI star site flanking preferences are distinct from those of cognate sites and that each asymmetric star half site may have distinct flanking preferences as well.

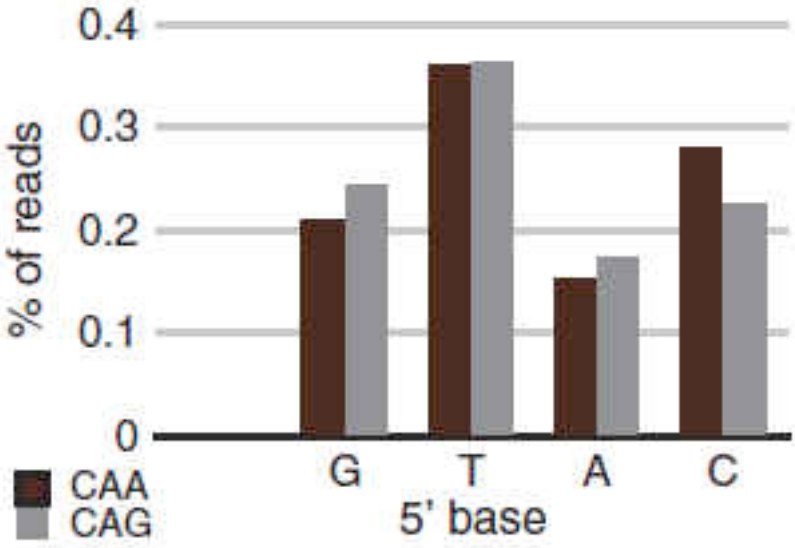

Figure 3. MfeI activity is affected by flanking base preference at CAACTG star sites. Bars represent the percentage of wild-type MfeI star activity assay reads mapping to CAACTG sites having a particular 50 adjacent base, with a higher percentage indicating that adjacent base creates a favourable context for digestion. Because the star site is asymmetric, adjacent base preferences are shown for the two half sites, CAA (black) and CAG (gray).

DISCUSSION

The power of next-generation sequencing has typically been applied to the characterization of the sequence or function of a genome. Here we use the massively parallel nature of next-generation sequencing to assay the enzymatic activity of restriction endonucleases that cleave both strands of double-stranded DNA. We developed a novel assay to allow the characterization of restriction enzyme recognition sites without any prior knowledge, and also used the related RAD-Seq method to assay the effect of flanking sequence on restriction enzyme cleavage.

We first quantitatively assayed the activity of both EcoRI and MfeI and their high-fidelity counterparts (EcoRI-HF and MfeI-HF) by mapping cleavage events to the *E. coli* reference genome. For each enzyme the majority of reads mapped to the cognate sites, demonstrating the correlation between cleavage efficiency and read count as well as highlighting the method's utility in *de novo* recognition site discovery. This unbiased detection method also simultaneously quantified star activity over all DNA configurations present in the *E. coli* genome. The star activity occurred at sites with one base pair substitutions with respect to the cognate sites as has been previously observed[19]. For both enzymes only a subset of the possible single substitution sites produced sequence reads, which effectively identified those degenerate sites capable of generating appreciable star activity. Different star sites showed a wide range of activity indicating the degree to which specific base changes are tolerated by the restriction enzyme. In the case of EcoRI, the three most abundant star sites in our data (GAATTT, GAAGTC and GAATTA) have been previously shown to be the three most efficiently cleaved[17,18]. The high-fidelity restriction enzymes developed by New England Biolabs showed drastically

reduced star activity compared to wild type. The assay was able to quantify this reduction across all potential cleavage sites and validate that no major cryptic DNA sequences are cleaved by the engineered high fidelity variants.

Massively parallel sequencing was also used to quantify the flanking preferences of MfeI. The relative presence of flanking nucleotides in sequence reads generated from the same complex substrate was compared across 12 enzyme concentrations using RAD-Seq. *Drosophila* genomic DNA was used as substrate in order to provide sufficient diversity of MfeI sites. Under enzyme saturation, an equal contribution of reads from sites was observed regardless of flanking sequences. As enzyme concentration was decreased, flanking nucleotide preferences of progressively larger magnitude were observed. When reads were binned by a single flanking nucleotide, **G**-CAATTG sites were shown to be favorable while **A**-CAATTG and **C**-CAATTG were shown to be unfavorable and **T**-CAATTG was relatively neutral. Binning reads by flanking dinucleotides showed even larger effects. While the general trends seen when examining single flanking nucleotides were still apparent, the dinucleotide analysis underscored the unique energetic contributions to cleavage of each unique sequence context. This was shown in our data by the ability of a given nucleotide in the second position away from the cut site to confer either a positive or negative effect on cleavage depending on the identity of the adjacent nucleotide. For example, the thymine nucleotide in the second position away from the cut site led to increased cleavage for **T**G-CAATTG, **T**A-CAATTG and **T**T-CAATTG but decreased cleavage for **T**C-CAATTG.

We also analyzed MfeI star activity from the first set of experiments with respect to flanking sequence. While the *E. coli* genome is not of sufficient complexity for

exhaustive flanking sequence analysis, it does provide enough diversity to confidently investigate effect of the adjacent base. Because digestion at the CAACTG star site was incomplete, we expected flanking preferences to be apparent much as they were in enzyme-limiting conditions at cognate MfeI sites. Indeed, we found that the flanking sequence affected CAACTG star site cleavage as well. In contrast to the palindromic MfeI cognate site, flanking preferences differed on each side of the asymmetric star sites. Notably, the MfeI star site flanking preferences are distinct from the cognate site flanking preferences, which is consistent with biophysical work suggesting star site-enzyme complexes are profoundly different from their cognate counterparts[2,34].

In this paper we present new high-throughput methods to characterize restriction endonuclease activity. The two techniques link Illumina sequence reads to cleavage events of highly complex substrate provided by sequenced genomes to assay enzyme activity in a highly parallel fashion. The data acquired from their application to MfeI and EcoRI is consistent with previously described principles regarding restriction enzyme activity. These techniques are easily applied to both previously characterized and newly discovered type II restriction endonucleases. Genome sequencing has yielded many thousands of putative restriction endonucleases[26], so the ability to quickly characterize their activity over all possible recognition sites will yield novel target specificities at a much higher rate than is currently possible. Additionally, the structure-function relationship of restriction enzymes has been long-studied; these methods provide a rapid way to generate data about target specificity and activity for enzymes in altered conditions or altered protein structure. Thus, the methodology presented and validated in

this study will serve as a basis for applying the power of massively parallel analysis to the active and essential field of restriction enzymology.

BRIDGE TO CHAPTER V

The work described in the following chapter directs the power of high-throughput sequencing toward gene regulation in a new way. Like the restriction endonuclease assay described above, sequencing reads are linked through molecular library preparation to biological events. In this case, transcriptional activation from a putative enhancer region is linked to output of a transcribed barcode that can be counted in high-throughput sequencing experiments. This allows for millions of stretches of DNA to be assayed in parallel for their ability to initiate transcription from a minimal promoter. We apply this new method to study hypoxia in *Drosophila* and couple the results with RNA-Seq data showing differentially expressed genes to provide a genome-wide analysis of the hypoxic regulatory network.

CHAPTER V

GENOME-WIDE IDENTIFICATION OF HYPOXIC ENHANCERS

I was the primary contributor to all aspects of the work with experimental design and analysis contributions from Johnson EA and experimental contributions from Preston JL and Randel MA.

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 subunit[6]. 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-1[9,10], 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 400-500 bp 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 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 mRNAs[20], 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 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 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

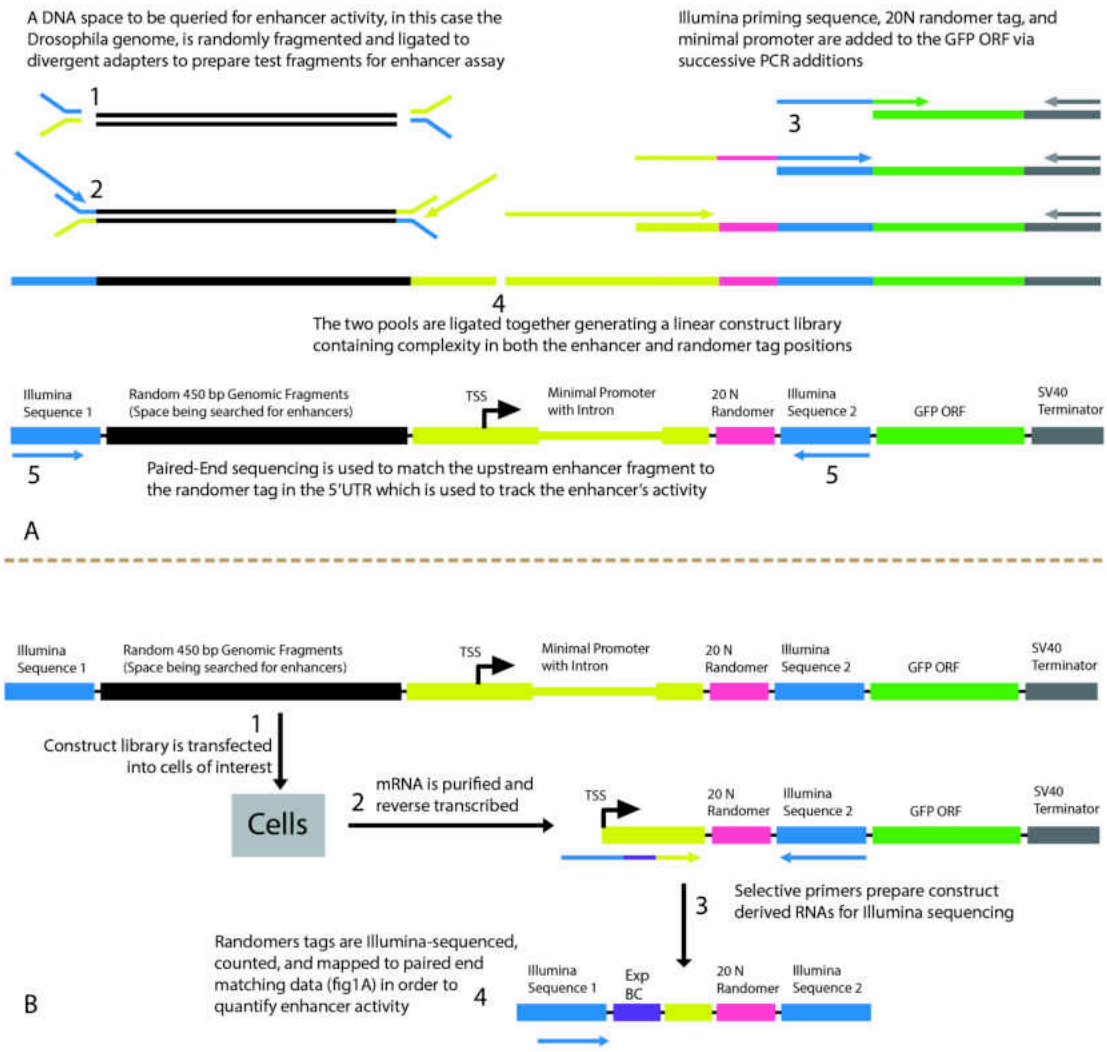

Figure 1. Enhancer library synthesis and assay. (A) DNA of interest is fragmented (step 1) and ligated to divergent adapters (step 2) leaving potential enhancer fragments with Illumina sequence on one side and the beginning of the synthetic minimal promoter on the other. The GFP gene is used as a template 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.

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. 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 (1819-1919 bp) 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 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 TopHat[22]. 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

Bowtie2[23] 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 Bowtie2[23]. 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 DESeq[24] 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 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

Table 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].

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

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].

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 2, Table 1). All six of these are the homologous Hsp70B enhancers.

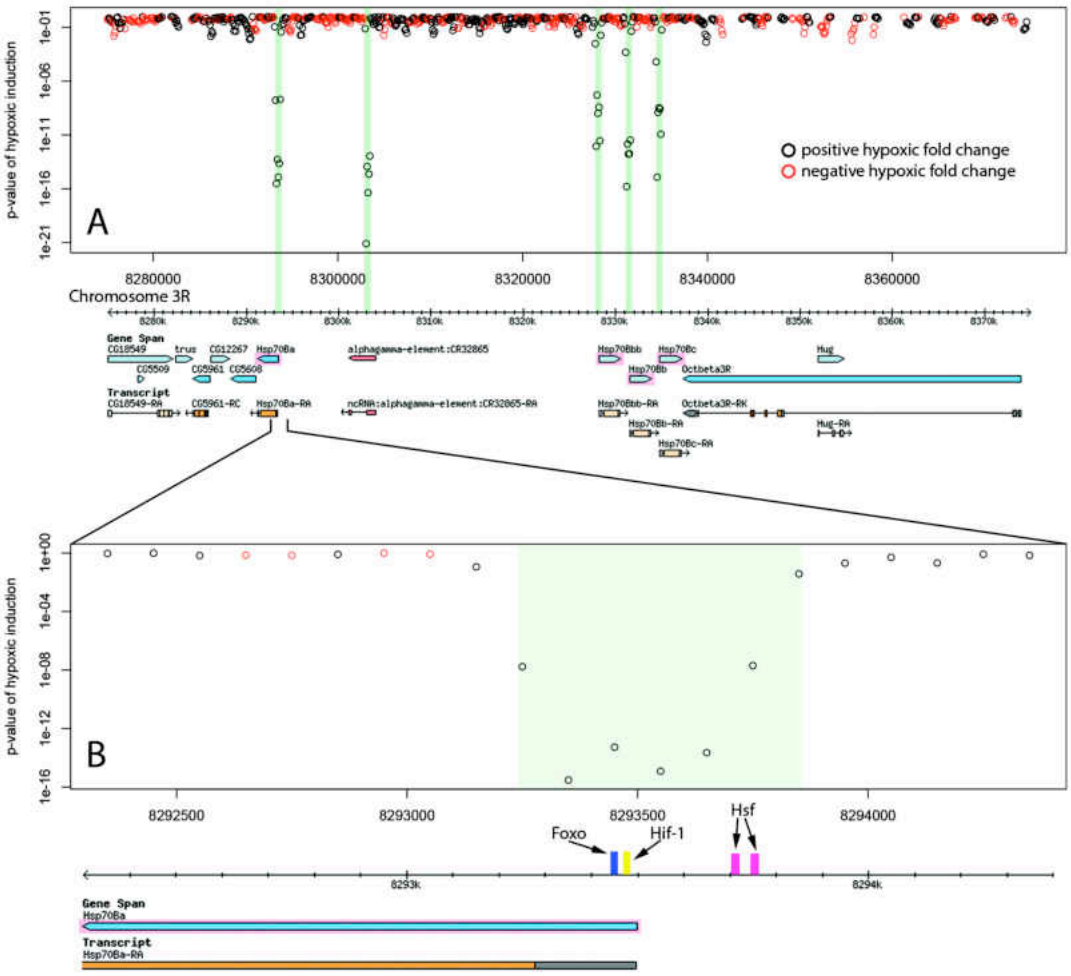

Figure 2. Hypoxic enhancer activity by 100bp bins at the Hsp70B locus. Each open circle plots the p-value of the difference in randomer tag counts mapping to that 100bp bin between normoxia and hypoxia. Green bars show enhancer regions discovered by our genome-wide screen. (A) The four Hsp70B homologues 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 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).

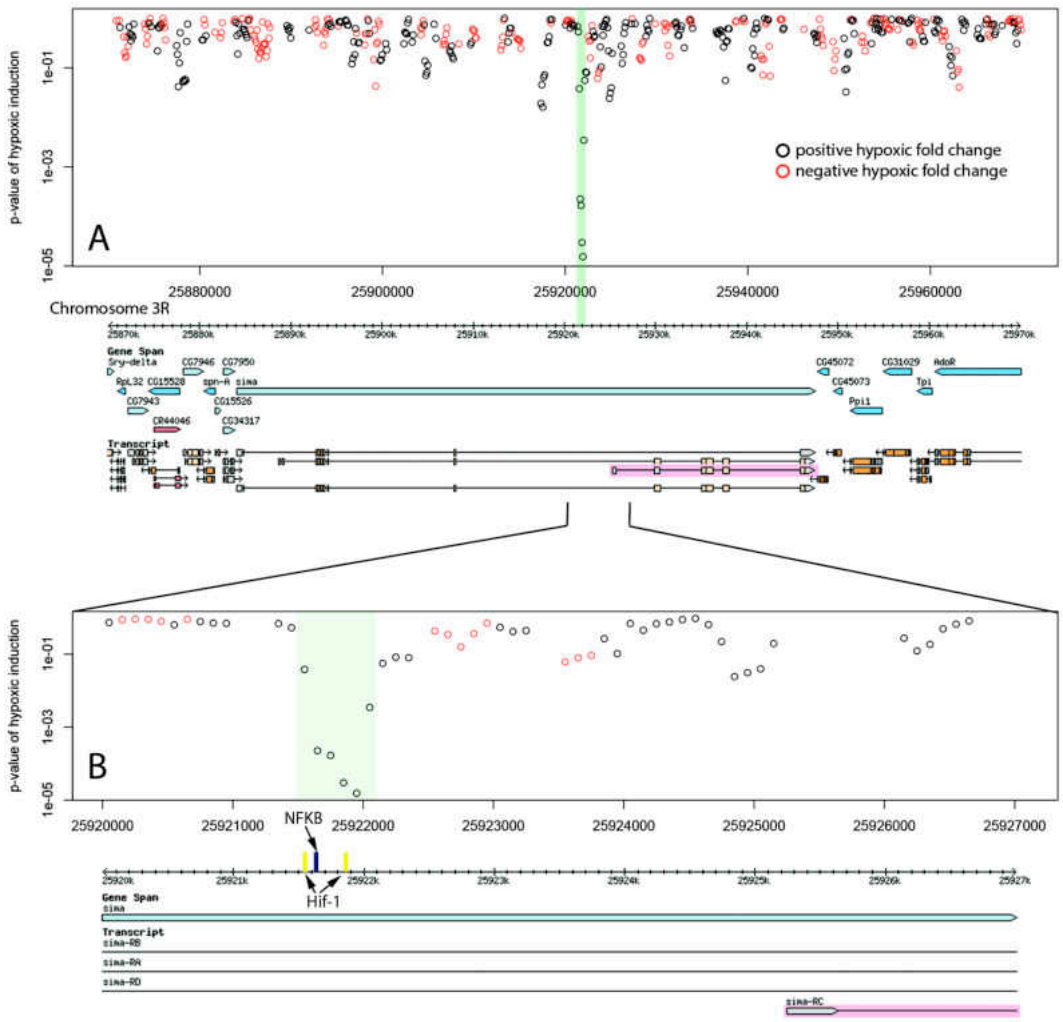

Figure 3. Hypoxic enhancer activity by 100bp bins at the Sima (HIF-1α) locus. Each open circle plots the p-value of the difference in randomer tag counts mapping to that 100bp bin between normoxia and hypoxia. The green bar shows the enhancer region discovered by our genome-wide screen. (A) HIF-1 is 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 4).

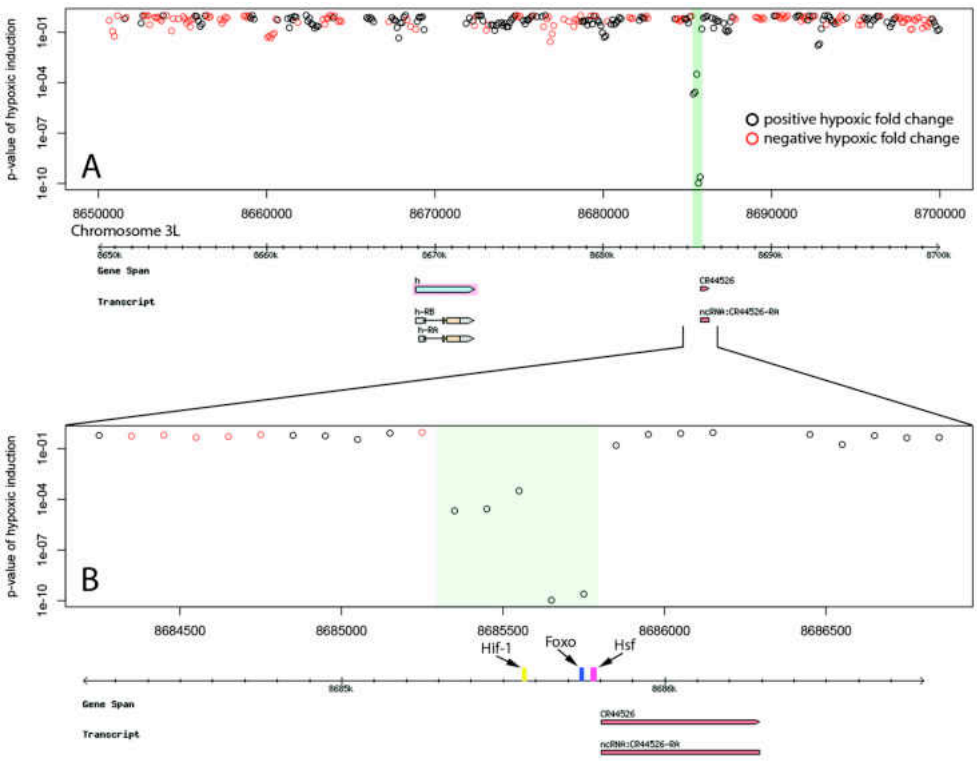

Figure 4. Hypoxic enhancer activity by 100bp bins at the hairy locus. Each open circle plots the p-value of the difference in randomer tag counts mapping to that 100bp bin between normoxia and hypoxia. The green bar shows the enhancer region discovered by our genome-wide screen. (A) The hairy gene produces 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 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). Binding sites occurring in each individual enhancer are listed in Table 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 2). The intronic Sima enhancer (Figure 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 4).

Table 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.

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

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

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 1). Enhancer regions in intergenic space corresponded with known ncRNA loci and in each case the ncRNA was itself up-regulated under hypoxia (Table 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.

CHAPTER VI

CONCLUSIONS

Genomics has become a foundational discipline whose influence has touched all aspects of biological research. The ability to efficiently identify and contextualize causal nucleic acid stretches has opened new avenues in systems biology, developmental biology, and population genetics. Additionally, the genomics era has brought exciting progress to health science as genome-wide studies provide new insights into regenerative and personalized medicine. Synthetic biology has also advanced greatly with next-generation genomic techniques allowing massively parallel quantification of regulatory modules. The work presented in this dissertation employs and develops massively parallel genomics techniques to identify and contextualize functional DNA loci across species and applications.

In Chapter II we looked into the genome of the model organism *C. elegans* using RAD-Seq. This long studied roundworm has played an important role in identifying conserved metazoan genes involved in basic cellular processes. We advanced this pursuit by employing RAD-Seq to rapidly identify genes essential to mitotic spindle assembly. Chapter III directs RAD-Seq to query the genomes of Montana salmonids in the North Fork of the Flathead River. Here genomics transcends the research arena to inform natural resource management. The large number of population-specific DNA polymorphisms identified by RAD-Seq allowed deep analysis of the impact of introduced rainbow trout on the indigenous cutthroat populations. Indeed, evidence of introgression

was discovered in our genomics-era approach that eluded previous low-throughput DNA analyses.

Chapter IV moves from association studies towards functional applications of massively-parallel sequencing. We presented a novel assay capable of quantifying the cleavage activity of restriction enzymes over millions of DNA sequences simultaneously. The assay parlays the existence of sequenced genomic DNA as a complex experimental substrate. We linked sequencing events to restriction digest events in order to quantify the specificity of EcoRI and MfeI across all possible sites in parallel. This yielded valuable information about their off-target activity as well as context-specific effects that would have been drastically more difficult to obtain using previous methods. The specific details about MfeI and EcoRI activity will be of use in molecular cloning and the presented technique can be used to characterize the large number of restriction enzymes recognized in bacterial metagenomic sequence.

Chapter V applied genomics tools to study live cells. We developed a technique capable of testing millions of potential DNA regulatory elements in parallel inside cultured insect cells and used it to investigate the hypoxic regulatory network. By using random barcodes to track the transcriptional activity of fragments spanning the *Drosophila* genome, we were able to provide the first genome-wide screen of regulatory elements mediating the response to an environmental stress. We then analyzed this data with respect to endogenously upregulated genes seen in our RNA-Seq data and transcription factor binding motifs. We uncovered an exciting level of complexity as enhancer elements often fell far outside of promoter regions and often integrated signals from multiple stress response pathways. Many were even associated with differentially

expressed ncRNAs. This study provided an empirical set of DNA sequences capable of enhancing hypoxic transcription and a generalizable method for transcriptional enhancer discovery. Taken together, the work presented in this dissertation makes significant progress in a variety of biological research areas unified by genomic theory and methodology.

REFERENCES CITED

Chapter I

1. Schena, M et al. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." *Science* 270 (1995): 467-470.

2. Heller, RA et al. "Discovery and analysis of inflammatory disease-related genes using cDNA microarrays." *Proceedings of the National Academy of Sciences* 94 (1997): 2150-2155.

3. Debouck, C and Goodfellow, PN. "DNA microarrays in drug discovery and development." *Nature genetics* 21 (1999): 48-50.

4. Shinozaki K et al. "Regulatory network of gene expression in the drought and cold stress responses." *Current opinion in plant biology* 6 (2003): 410-417.

5. Khan, J et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." *Nature medicine* 7 (2001): 673-679.

6. Miller, MR et al. "Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers." *Genome research* 17 (2007): 240-248.

7. Bennett, S. "Solexa ltd." *Pharmacogenomics* 5 (2004): 433-438.

8. Wheeler, DA et al. "The complete genome of an individual by massively parallel DNA sequencing." *Nature* 452 (2008): 872-876.

9. Zerbino, DR and Birney, E. "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." *Genome research* 18 (2008): 821-829.

10. Huang, Y. "The duck genome and transcriptome provide insight into an avian influenza virus reservoir species." *Nature genetics* 45 (2013): 776-783.

11. Groenen, M et al. "Analyses of pig genomes provide insight into porcine demography and evolution." *Nature* 491 (2012): 393-398.

12. Van Bakel H et al. "The draft genome and transcriptome of Cannabis sativa." *Genome biology* 12 (2011): R102.

13. Baird, NA et al. "Rapid SNP discovery and genetic mapping using sequenced RAD markers." *PloS One* 3 (2008): e3376.

14. Mortazavi, A et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nature methods* 5 (2008): 621-628.

15. Valouev, A et al. "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data." *Nature methods* 5 (2008): 829-834.

16. Wang, Z et al. "RNA-Seq: a revolutionary tool for transcriptomics." *Nature Reviews Genetics* 10 (2009): 57-63.

17. Pickrell, JK et al. "Understanding mechanisms underlying human gene expression variation with RNA sequencing." *Nature* 464 (2010): 768-772.

18. Trapnell, C et al. "TopHat: discovering splice junctions with RNA-Seq." *Bioinformatics* 25 (2009): 1105-1111.

19. Wang, E et al. "Alternative isoform regulation in human tissue transcriptomes." *Nature* 456 (2008): 470-476.

20. Friedländer, MR et al. "Discovering microRNAs from deep sequencing data using miRDeep." *Nature biotechnology* 26 (2008): 407-415.

21. Andersson, R et al. (2014) An atlas of active enhancers across human cell types and tissues. Nature 507: 455-461.

22. Kapranov, P et al. "New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism." *Nature* 466 (2010): 642-646.

23. Song, L et al. "Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity." *Genome research* 21 (2011): 1757-1767.

24. Patwardhan RP, et al. “Massively parallel functional dissection of mammalian enhancers in vivo.” *Nature Biotechnology* 30 (2012): 265-270.

25. Arnold, CD et al. “Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq.” *Science* 339 (2013): 1074-1077.

Chapter II

1. Sarin, S et al. “Caenorhabditis elegans mutant allele identification by whole-genome sequencing.” *Nature Methods* 5 (2008): 865–867.

2. Smith, D et al. “Rapid whole-genome mutational profiling using nextgeneration sequencing technologies.” *Genome Research* 18 (2008): 1638–1642.

3. Srivatsan, A et al. “High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies.” *PLoS Genetics* 4 (2008): e1000139.

4. Blumenstiel, JP et al. “Identification of EMS-induced mutations in Drosophila melanogaster by whole-genome sequencing.” *Genetics* 182 (2009): 25–32.

5. Irvine, D et al. “Mapping epigenetic mutations in fission yeast using whole-genome next-generation sequencing.” *Genome Research* 19 (2009): 1077–1083.

6. Doitsidou, M et al. “C. elegans mutant identification with a one-step wholegenome-sequencing and SNP mapping strategy.” *PLoS ONE* 5 (2010): e15435.

7. Zuryn, S et al. “A strategy for direct mapping and identification of mutations by wholegenome sequencing.” *Genetics* 186 (2010): 427–430.

8. Ksiazek, TG et al. “A novel coronavirus associated with severe acute respiratory syndrome.” *New England Journal of Medicine* 348 (2003): 1953–1966.

9. Rota, PA et al. “Characterization of a novel coronavirus associated with severe acute respiratory syndrome.” *Science* 300 (2003): 1394– 1399.

10. Wang, D et al. “Viral discovery and sequence recovery using DNA microarrays.” *PLoS Biology* 1 (2003): E2.

11. Albert, TJ et al. “Direct selection of human genomic loci by microarray hybridization.” *Nature Methods* 4 (2007): 903–905.

12. Hodges, E et al. “Genome-wide in situ exon capture for selective resequencing.” *Nature Genetics* 39 (2007): 1522–1527.

13. Okou, DT et al. “Microarray-based genomic selection for highthroughput resequencing.” *Nature Methods* 4 (2007): 907–909.

14. Gnirke, A et al. “Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.” *Nature Biotechnology* 27 (2009): 182–189.

15. Brenner, S. “The genetics of Caenorhabditis elegans.” Genetics 77 (1974): 71–94.
16. Encalada, SE et al. “DNA replication defects delay cell division and disrupt cell polarity in early Caenorhabditis elegans embryos.” *Developmental Biology* 228 (2000): 225–238.

17. Baird, NA et al. “Rapid SNP discovery and genetic mapping using sequenced RAD markers.” *PLoS ONE* 3 (2008): e3376.

18. Hohenlohe, PA et al. “Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags.” *PLoS Genetics* 6 (2010): e1000862.

19. Langmead, B et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.” *Genome Biology* 10 (2009): R25.

20. Li, H et al. “The Sequence Alignment/Map format and SAMtools.” *Bioinformatics* 25 (2009): 2078–2079.

21. Bigelow, H et al. “MAQGene: software to facilitate C. elegans mutant genome sequence analysis.” *Nature Methods* 6 (2009): 549.

22. Lewis, ZA et al. “High-density detection of restriction-site-associated DNA markers for rapid mapping of mutated loci in Neurospora.” *Genetics* 177 (2007): 1163–1171.

23. Miller, MR “Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers.” *Genome Research* 17 (2007): 240–248.

24. Coyne, KJ et al. “Modified serial analysis of gene expression method for construction of gene expression profiles of microbial eukaryotic species.” *Applied Environmental Microbiology* 70 (2004): 5298– 5304.

25. Sonnichsen, B et al. “Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans.” *Nature* 434 (2005): 462–469.

26. Harris, TW et al. “WormBase: a comprehensive resource for nematode research.” *Nucleic Acids Research*. 38 (2010): D463–D467.

27. Kemphues, KJ et al. “Maternal-effect lethal mutations on linkage group II of Caenorhabditis elegans.” *Genetics* 120 (1988): 977–986.

28. Jorgensen, EM and Mango, SE. “The art and design of genetic screens: Caenorhabditis elegans.” Nature Reviews Genettics. 3 (2002): 356–369.

29. O’Rourke, SM et al. “A survey of new temperature-sensitive, embryonic lethal mutations in C. elegans: 24 alleles of thirteen genes.” *PLoS ONE* 6 (2011): e16644.

30. O’Rourke, SM et al. “Dynein modifiers in C. elegans: light chains suppress conditional heavy chain mutants.” *PLoS Genetics* 3 (2007): e128.

31. Baxter, SW et al. “Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism.” *PLoS One* 6 (2011): e19315.

32. Chutimanitsakun, Y et al. “Construction and application for QTL analysis of a restriction site associated DNA (RAD) linkage map in barley.” *BMC Genomics* 12 (2011):4.

33. Johnsen, RC and Baillie, DL. “Genetic analysis of a major segment of the genome of Caenorhabditis elegans.” *Genetics* 129 (1991): 735–752.

34. Clark, DV and Baillie, DL. “Genetic analysis and complementation by germ-line transformation of lethal mutations in the unc-22 IV region of Caenorhabditis elegans.” *Molecular Genetics and Genomics* 232 (1992): 97–105.

35. Stewart, HI et al. “Lethal mutations defining 112 complementation groups in a 4.5 Mb sequenced region of Caenorhabditis elegans chromosome III.” *Molecular Genetics and Genomics* 260 (1998): 280–288.

36. Johnsen, RC et al. “Mutational accessibility of essential genes on chromosome I in Caenorhabditis elegans.” *Molecular Genetics and Genomics*. 263 (2000): 239–252.

37. Greenwald, IS and Horvitz, HR. “unc-93(e1500): a behavioral mutant of Caenorhabditis elegans that defines a gene with a wild-type null phenotype.” *Genetics* 96 (1980): 147–164.

Chapter III

1. Allendorf, FW et al. “The problems with hybrids: setting conservation guidelines.” *Trends in Ecology and Evolution* 16 (2001): 613–622.

2. Muhlfeld, CC et al. “Hybridization rapidly reduces fitness of a native trout in the wild.” *Biology Letters* 5 (2009): 328–331.

3. Muhlfeld, CC et al. “Spatial and temporal spawning dynamics of native westslope cutthroat trout, Oncorhynchus clarkii lewisi, introduced rainbow trout, Oncorhynchus mykiss, and their hybrids.” *Canadian Journal of Fisheries and Aquatic Sciences* 66 (2009): 1153–1168.

4. Kelly, BP et al. “The Arctic melting pot.” *Nature*, 468 (2010): 891.

5. Fitzpatrick, BM et al “Rapid spread of invasive genes into a threatened native species.” *Proceedings of the National Academy of Sciences USA* 107 (2010): 3606–3610.

6. Gompert, Z and Buerkle, CA. “A powerful regression-based method for admixture mapping of isolation across the genome of hybrids.” *Molecular Ecology*, 18 (2009): 1207–1224.

7. Miller, JM et al. “Genomic consequences of genetic rescue in an insular population of bighorn sheep (Ovis canadensis).” *Molecular Ecology* 21 (2012): 1583–1596.

8. Teeter KC et al. “The variable genomic architecture of isolation between hybridizing species of house mice.” *Evolution* 64 (2010): 472–485.

9. Halverson, A. “An Entirely Synthetic Fish: How Rainbow Trout Beguiled America and Overran the World.” *Yale University Press* (2010).

10. Shepard, BB et al. “Status and conservation of westslope cutthroat within the western United States.” *North American Journal of Fisheries Management* 25 (2005): 1426–1440.

11. Hitt, NP et al. “Spread of hybridization between native westslope cutthroat trout, Oncorhynchus clarki lewisi, and nonnative rainbow trout, Oncorhynchus mykiss.” *Canadian Journal of Fisheries and Aquatic Sciences* 60 (2003): 1440–1451.

12. Boyer, MC et al. “Rainbow trout (Oncorhynchus mykiss) invasion and the spread of hybridization with native westslope cutthroat trout (Oncorhynchus clarkia lewisii).” *Canadian Journal of Fisheries and Aquatic Sciences* 65 (2008): 658–669.

13. Muhlfeld CC et al. “Local habitat, watershed, and biotic factors influencing the spread of hybridizations between native westslope cutthroat trout and introduced rainbow trout.” *Transactions of the American Fisheries Society* 138 (2009): 1036–1051.

14. Perry, GML et al. “Quantitative trait loci for upper thermal tolerance in outbred strains of rainbowtrout (Oncorhynchus mykiss).” *Heredity* 86 (2001): 333–341.

15. Narum, SR et al. “Adaptation of redband trout in desert and montane environments.” *Molecular Ecology* 19 (2010): 4622–4637.

16. Angeloni, F et al. “Genomic toolboxes for conservation biologists.” *Evolutionary Applications* 5 (2011): 130–143.

17. Morin, PA et al. “SNPs in ecology, evolution and conservation.” *Trends in Ecology and Evolution* 19 (2004): 208–216.

18. Seeb, JE et al. “SNP genotyping by the 5′-nuclease reaction: advances in high throughput genotyping with non-model organisms.” *Methods in Molecular Biology, Single Nucleotide Polymorphisms,* 2nd edn (2009): 277–292.

19. Twyford, AD and Ennos, RA “Next-generation hybridization and introgression.” *Heredity* 108 (2012): 179–189.

20. Finger, AJ et al. Six diagnostic single nucleotide polymorphism markers for detecting introgression between cutthroat and rainbow trout. *Molecular Ecology Resources* 9 (2009): 759–763.

21. Harwood, AS and Phillips, RB “A suite of twelve single nucleotide polymorphism markers for detecting introgression between cutthroat and rainbow trout. *Molecular Ecology Resources* 11 (2011): 382–385.

22. McGlauflin, MT et al. “High-resolution melting analysis for the discovery of novel singlenucleotide polymorphisms in rainbow and cutthroat trout for species identification.” *Transactions of the American Fisheries Society* 139 (2010): 676–684.

23. Amish, SJ et al. “RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays.” *Molecular Ecology Resources* 12 (2012): 653–660.

24. Campbell, NR et al. “Development and evaluation of 200 novel SNP assays for population genetic studies of westslope cutthroat trout and genetic identification of related taxa.” *Molecular Ecology Resources* 12(2012): 942–949.

25. Pritchard, VL et al. “Discovery and characterization of a large number of diagnostic markers to discriminate Onchorhynchus mykiss and O. clarkii.” *Molecular Ecology Resources* 12 (2012): 918–931.

26. Allendorf, FW and Danzmann, RG. “Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout.” *Genetics* 145 (1997): 1083–1092.

27. Everett, MV et al. “Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome.” *Molecular Ecology Resources* 11 (2011): 93– 108.

28. Seeb, JE et al. “Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms.” *Molecular Ecology Resources* 11 (2011): 1–8.

29. Baird, NA et al. “Rapid SNP discovery and genetic mapping using sequenced RAD markers.” *PLoS One* 3 (2008): e3376.

30. Hohenlohe, PA et al. “Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout.” *Molecular Ecology Resources* 11 (2011): 117–122.

31. Davey, JW et al. “Genome-wide genetic marker discovery and genotyping using next-generation sequencing.” *Nature Reviews Genetics* 12 (2011): 499–510.

32. Baxter, SW et al. “Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism.” *PLoS One* 6 (2011): e19315.

33. Etter, PD et al. “Local de novo assembly of RAD paired-end contigs using short sequencing reads.” *PLoS One* 6 (2011) e18561.

34. Willing, EM et al “Paired-end RAD-seq for de-novo assembly and marker design without available reference.” *Bioinformatics* 27(2011): 2187–2193.

35. Etter, PD and Johnson, EA. “RAD paired-end sequencing for local de novo assembly and SNP discovery in non-model organisms.” *Data Production and Analysis in Population Genomics: Methods and Protocols* (2012): 135–151.

36. Etter, PD et al. “SNP discovery and genotyping for evolutionary genetics using RAD sequencing.” *Molecular Methods for Evolutionary Genetics* (2011): 157–178.

37. Catchen, JM et al. “Stacks: building and genotyping loci de novo from short-read sequences.” *G3 Genes Genomes Genetics* 1 (2011): 171–182.

38. Miller, MR et al. “A conserved haplotype controls parallel adaptation in geographically distant salmonid populations.” *Molecular Ecology* 21 (2011): 237–249.

39. Zerbino, DR and Birney, E. “Velvet: algorithms for de novo short read assembly using de Bruijn graphs.” *Genome Research* 18 (2008): 821–829.

40. Huang, X and Madan, A “CAP3: a DNA sequence assembly program.” *Genome Research* 9 (1999): 868–877.

41. Langmead, B et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.” *Genome Biology* 10 (2009): R25.

42. Hohenlohe, PA et al. (2010) “Population genomic analysis of parallel adaptation in threespine stickleback using sequenced RAD tags.” *PLoS Genetics* 6 (2010): e1000862.

43. Long, JC. (1991) “The genetic structure of admixed populations.” *Genetics*, 127 (1991): 417–428.

44. Fitzpatrick, BM et al. “Rapid fixation of non-native alleles revealed by genome-wide SNP analysis of hybrid tiger salamanders.” *BMC Evolutionary Biology* 9 (2009): 176.

45. Benjamini, Y and Hochberg, Y. “Controlling the false discovery rate: a practical and powerful approach to multiple testing.” *Journal of the Royal Statistical Society Series B*, 57 (1995): 289–300.

46. Davey JW et al. “Special features of RAD sequencing data: implications for genotyping.” *Molecular Ecology* 22 (2012): 3151–3164.

47. Ogden, R. “Unlocking the potential of genomic technologies for wildlife forensics.” *Molecular Ecology Resources* 11 (2011): 109–116.

48. Seeb, LW et al. “Single nucleotide polymorphisms across a species’ range implications for conservation studies of Pacific salmon.” *Molecular Ecology Resources* 11 (2011): 195–217.

49. Gompert, Z and Buerkle, CA. “A hierarchical Bayesian model for next-generation population genomics.” *Genetics* (187): 903– 917.

50. Buerkle, CA et al. “The n = 1 constraint in population genomics.” *Molecular Ecology* 20 (2011): 1575–1581.

51. Sunnucks, P. “Efficient genetic markers for population biology.” *Trends in Ecology and Evolution* 15 (2000): 199–203.

52. Beaumont, M and Rannala, B. “The Bayesian revolution in genetics.” *Nature Reviews Genetics* 5 (2004): 251–261.

53. Lamaze, FC et al. “Dynamics of introgressive hybridization assessed by SNP population genomics of coding genes in stocked brook charr (Salvelinus fontinalis).” *Molecular Ecology* 21 (2012): 2877–2895.

54. Schwartz, MK et al. “Genetic monitoring as a promising tool for conservation and management.” *Trends in Ecology and Evolution* 22 (2007): 25–33.

55. Luikart, G et al. “The power and promise of population genomics: from genotyping to genome typing.” *Nature Reviews Genetics* 4 (2003): 981–994.

56. Shine, R. “Invasive species as drivers of evolutionary change: cane toads in tropical Australia.” *Evolutionary Applications* 5 (2011): 107–116.

57. Moren, A et al. “Identification and characterization of LTBP-2, a novel latent transforming growth factor-b-binding protein.” *Journal of Biological Chemistry* 269 (1994): 32469–32478.

58. Oklu, R and Hesketh, R. “The latent transforming growth factor b binding protein (LTBP) family.” *Biochemical Journal* 352 (2000): 601–610.

59. Kosova, G et al. “Genome-wide association study identifies candidate genes for male fertility traits in humans.” *The American Journal of Human Genetics* 90 (2012): 950–961.

60. Hirani, R et al. “LTBP-2 specifically interacts with the amino-terminal region of fibrillin-2 and competes with LTBP-1 for binding to this microfibrillar protein.” *Matrix Biology* 26 (2007): 213–223.

61. Andersson, ML and Eggen, RI. “Transcription of the fish latent TGFbeta-binding protein gene is controlled by estrogen receptor alpha.” *Toxicology in vitro* 20 (2006): 417–425.

62. Lankford, SE and Weber, GM. “Temporal mRNA expression of transforming growth factor-beta superfamily members and inhibitors in the developing rainbow trout ovary.” *General and Comparative Endocrinology* 166 (2010): 250–258.

63. Gahr, SA et al. “Identification and expression of Smads associated with TGF-b/activin/nodal signaling pathways in the rainbow trout (Oncorhynchus mykiss).” *Fish Physiology and Biochemistry* 38 (2012): 1233–1244.

Chapter IV

1. Orlowski, J and Bujnicki JM. “Structural and evolutionary classification of Type II restriction enzymes based on theoretical and experimental analyses.” *Nucleic Acids Research* 36 (2008): 3552-3569.

2. Jen-Jacobson L. “Protein-DNA recognition complexes: conservation of structure and binding energy in the transition state.” *Biopolymers* 44 (1997): 153-180.

3. Wei, H et al. “The Fidelity Index provides a systematic quantitation of star activity of DNA restriction endonucleases.” *Nucleic Acids Research* 36 (2008): e50.

4. Kolesnikov, VA et al. “Relaxed specificity of endonuclease BamH1 as determined by identification or recognition sites in SV40 and pBR322 DNAs.” *FEBS Letters* 132 (1981):101-104.

5. Zylicz-Stachula, A et al. “Chemically-induced affinity star restriction specificity: a novel TspGWI/sinefungin endonuclease with theoretical 3-bp cleavage frequency.” *BioTechniques* 50 (2011): 397-406.

6. Saravanan, M et al. “Evolution of sequence specificity in a restriction endonuclease by a point mutation.” *Proceedings of the National Academy of Science USA* 105 (2008): 10344-10347.

7. Malyguine, E et al. “Alteration of the specificity of restriction endonucleases in the presence of organic solvents.” *Gene* 8 (1980): 163–177.

8. Alves, J et al. “The influence of sequences adjacent to the recognition site on the cleavage of oligodeoxynucleotides by the EcoRI endonuclease.” *European Journal of Biochemistry* 140 (1984): 83-92.

9. Wolfes, H et al. “A comparison of the structural requirements for DNA cleavage by the isoschizomers HaeIII, BspRI and BsuRI.” *European Journal of Biochemistry* 150 (1985): 105-110.

10. Horton NC and Perona JJ. “Recognition of flanking DNA sequences by EcoRV endonuclease involves alternative patterns of water-mediated contacts.” *Journal of Biological Chemistry* 273 (1998): 21721-21729.

11. Engler, LE et al. (2001) “The energetics of the interaction of BamHI endonuclease with its recognition site GGATCC.” *Journal of Molecular Biology* 307 (2001): 619-636.

12. Jen-Jacobson, L et al. “Thermodynamic Parameters of Specific and Nonspecific Protein-DNA Binding.” *Supramolecular Chemistry* 12 (2000) 143-160.

13. Bujnicki, JM “Phylogeny of the restriction endonuclease-like superfamily inferred from comparison of protein structures.” *Journal of Molecular Evolution* 50 (2000): 39-44.

14. Engler, LE et al. “Specific Binding by EcoRV Endonuclease to its DNA Recognition Site GATATC.” *Journal of Molecular Biology* 269 (1997): 82-101.

15. Deibert, M et al. “Crystal structure of MunI restriction endonuclease in complex with cognate DNA at 1.7 A resolution.” *EMBO Journal* 18 (1999): 5805-5816.

16. Lesser, DR et al. “Facilitated distortion of the DNA site enhances EcoRI endonuclease-DNA recognition.” *Proceeding of the National Academy of Science USA* 90 (1993): 7548-7552.

17. Lesser, DR et al. “The energetic basis of specificity in the Eco RI endonuclease-DNA interaction.” *Science* 250 (1990): 776-786.

18. Thielking, V et al. “Accuracy of the EcoRI restriction endonuclease: binding and cleavage studies with oligodeoxynucleotide substrates containing degenerate recognition sequences.” *Biochemistry* 29 (1990): 4682-4691.

19. Sidorova NY and Rau DC. “Differences between EcoRI nonspecific and "star" sequence complexes revealed by osmotic stress.” *Biophysical Journal* 87 (2004): 2564-2576.

20. Horton, NC and Perona, JJ. “Role of protein-induced bending in the specificity of DNA recognition: crystal structure of EcoRV endonuclease complexed with d(AGAT) + d(ATCTT).” *Journal of Molecular Biology* 277 (1998): 779-787.

21. Nasri, M and Thomas, D. “Relaxation of recognition sequence of specific endonuclease HindIII.” *Nucleic Acids Research* 14 (1986): 811-821.

22. Samuelson, JC and Xu, SY. “Directed evolution of restriction endonuclease BstYI to achieve increased substrate specificity.” *Journal of Molecular Biology* 319 (2002): 673-683.

23. Jurenaite-Urbanaviciene, S et al. “Generation of DNA cleavage specificities of type II restriction endonucleases by reassortment of target recognition domains.” *Proceedings of the National Academy of Sciences USA* 104 (2007): 10358-10363.

24. Joshi, HK et al. “Alteration of sequence specificity of the type II restriction endonuclease HincII through an indirect readout mechanism.” *Journal of Biological Chemistry* 281 (2006): 23852-23869.

25. Guan, S et al. “Alteration of sequence specificity of the type IIS restriction endonuclease BtsI.” *PLoS One* 5 (2010): e11787.

26. Roberts, RJ et al. “REBASE-a database for DNA restriction and modification: enzymes, genes and genomes.” *Nucleic Acids Research* 38 (2010): D234-D236.

27. Xu, SY et al. “Discovery of natural nicking endonucleases Nb.BsrDI and Nb.BtsI and engineering of top-strand nicking variants from BsrDI and BtsI.” *Nucleic Acids Research* 35 (2007): 4608-4618.

28. Kostiuk, G et al. “Degenerate sequence recognition by the monomeric restriction enzyme: single mutation converts BcnI into a strand-specific nicking endonuclease.” *Nucleic Acids Research* 39 (2011): 3744-3753.

29. Xu,Y et al. “Engineering a nicking endonuclease N.AlwI by domain swapping.” *Proceedings of the National Academy of Science USA*, 98 (2001): 12990-12995.

30. Samuelson, JC et al. “Engineering a rare-cutting restriction enzyme: genetic screening and selection of NotI variants.” *Nucleic Acids Research* 34 (2006): 796-805.

31. Wang, H et al. “Comparative characterization of the PvuRts1I family of restriction enzymes and their application in mapping genomic 5-hydroxymethylcytosine.” *Nucleic Acids Research* 39 (2011): 9294-9305.

32. Johnson, DS et al. “Genome-wide mapping of in vivo protein-DNA interactions.” *Science* 316 (2007): 1497-1502.

33. Baird, NA et al. “Rapid SNP discovery and genetic mapping using sequenced RAD markers.” *PLoS One* 3 (2008): e3376.

Chapter V

1. Bulger M and Groudine M. “Functional and mechanistic diversity of distal transcription enhancers.” *Cell* 44 (2011): 327-339.

2. Perry, MW et al. “Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo.” *Proceedings of the Natlional Academy of Science USA,* 108 (2011): 13570-13575.

3. Lagha, M et al. “Mechanisms of transcriptional precision in animal development.” *Trends in Genetics* 28 (2012): 409-416.

4. Arnosti, DN and Kulkarni, MM. “Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?” *Journal of Cellular Biochemistry* 94 (2005): 890-898.

5. Swanson, C et al. “Rapid evolutionary rewiring of a structurally constrained eye enhancer.” *Current Biology* 21 (2011): 1186-1196.

6. Bruick, RK and McKnight, SL. “A conserved family of prolyl-4-hydroxylases that modify HIF.” *Science Signalling* 294 (2001): 1337-1340.

7. Lavista-Llanos S et al. “Control of the hypoxic response in Drosophila melanogaster by the basic helix-loop-helix PAS protein similar.” *Molecular and Cellular Biology* 22 (2002): 6842-6853.

8. Wang, GL and Semenza, GL. “General involvement of hypoxia-inducible factor 1 in transcriptional response to hypoxia.” *Proceedings of the National Academy of Science USA* 90 (1993): 4304–4308.

9. Rius, J et al. “NF-kappaB links innate immunity to the hypoxic response through transcriptional regulation of HIF-1alpha.” *Nature* 453 (2008): 807–811.

10. van Uden, P et al. “Evolutionary Conserved Regulation of HIF-1β by NF-κB.” *PLoS Genetics* 7 (2011): e1001285.

11. Scortegagna, M et al. “HIF-1α regulates epithelial inflammation by cell autonomous NFκB activation and paracrine stromal remodeling.” *Blood* 111 (2008): 3343-3354.

12. Metzker, ML. “Sequencing technologies—the next generation.” *Nature Reviews Genetics* 11 (2010): 31-46.

13. Wang, Z et al. “RNA-Seq: a revolutionary tool for transcriptomics.” *Nature Reviews Genetics* 10 (2009): 57-63.

14. Johnson, D et al. “Genome-wide mapping of in vivo protein-DNA interactions.” *Science* 316 (2007): 1497-1502.

15. Kwasnieski, JC et al. “Complex effects of nucleotide variants in a mammalian cis-regulatory element.” *Proceedings of the National Academy of Science USA*, 109 (2002): 19498-19503.

16. Patwardhan, RP et al. “Massively parallel functional dissection of mammalian enhancers in vivo.” *Nature Biotechnology* 30 (2012): 265-270.

17. Melnikov, A et al. “Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.” *Nature Biotechnology* 30 (2012): 271-277.

18. Kheradpour, P et al. “Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay.” *Genome Research* 23 (2013): 800-811.

19. Arnold, CD et al. “Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq.” *Science* 339 (2013): 1074-1077.

20. Zieler, H and Huynh, CQ. “Intron-dependent stimulation of marker gene expression in cultured insect cells.” *Insect Molecular Biology* 11 (2002): 87-95.

21. Pfeiffer, BD et al. “Tools for neuroanatomy and neurogenetics in Drosophila.” *Proceedings of the National Academy of Science USA* 105 (2008): 9715-9720.

22. Trapnell, C et al. “TopHat: discovering splice junctions with RNA-Seq.” *Bioinformatics* 25 (2009): 1105-1111.

23. Langmead, B and Salzberg, SL. “Fast gapped-read alignment with Bowtie 2.” *Nature Methods* 9 (2012): 357-359.

24. Anders, S and Huber, W. “Differential expression analysis for sequence count data.” *Genome Biology* 11 (2010): R106.

25. Ma, Q et al. “DMINDA: an integrated web server for DNA motif identification and analyses.” *Nucleic Acids Research* (2014): gku315.

26. Mathelier, A et al. “JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.” *Nucleic Acids Research* (2013): gkt997.

27. Li, Y et al. “HIF-and Non-HIF-Regulated Hypoxic Responses Require the Estrogen-Related Receptor in Drosophila melanogaster.” *PLoS Genetics* 9 (2013): e1003230.

28. Liu, G et al. “Identification and function of hypoxia-response genes in Drosophila melanogaster.” *Physiological Genomics,* 25 (2006): 134-141.

29. Zhou, D et al. “Mechanisms underlying hypoxia tolerance in Drosophila melanogaster: hairy as a metabolic switch.” *PLoS Genetics* 4 (2008): e1000221.

30. O'Keefe, LV et al. “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 (2011): 497-509.

31. Scuderi, A et al. “scylla and charybde, homologues of the human apoptotic gene RTP801, are required for head involution in Drosophila.” *Developmental Biology* 291 (2006): 110-122.

32. Tian, S et al. “Phylogeny disambiguates the evolution of heat-shock cis-regulatory elements in Drosophila.” *PLoS One* 5 (2010): e10669.

33. Li, D et al. “Isolation and functional analysis of the promoter of the amphioxus Hsp70a.” *Gene* 510 (2012): 39-46.

34. Hsu, AL et al. “Regulation of aging and age-related disease by DAF-16 and heat-shock factor.” *Science* 300 (2003): 1142-1145.

35. Wang, MC et al. “JNK extends life span and limits growth by antagonizing cellular and organism-wide responses to insulin signaling.” *Cell* 12 (2005): 115-125.

36. Acevedo, JM et al. “Oxygen Sensing in *Drosophila*: Multiple Isoforms of the Prolyl Hydroxylase Fatiga Have Different Capacity to Regulate HIFα/Sima.” *PLoS One* 5 (2010): e12390.

37. Andersson, R et al. “An atlas of active enhancers across human cell types and tissues.” *Nature* 507 (2014): 455-461.