

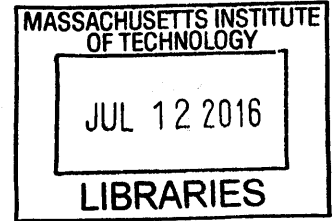
**Computational Personal Genomics:
understanding the functional effects of sequence
variation**

by

Robert C. Altshuler

Sc.B. Computer Science, Brown University (2001)

Sc.M. Computer Science, Brown University (2003)



Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2016 [June 2016]

© Robert C. Altshuler, MMXVI. All rights reserved.

The author hereby grants to MIT permission to reproduce and
distribute publicly paper and electronic copies of this thesis
document in whole or in part.

Signature redacted

Author

Department of Electrical Engineering and Computer Science

January 28, 2016

Signature redacted

Certified by


 Manolis Kellis

Professor of Computer Science

Thesis Supervisor

Signature redacted

Accepted by

 Leslie A. Kolodziejski

Chair of the Committee on Graduate Students

The author hereby grants to MIT permission to
reproduce and to distribute publicly paper and
electronic copies of this thesis document in
whole or in part in any medium now known or
hereafter created.

Computational Personal Genomics: understanding the functional effects of sequence variation

by

Robert C. Altshuler

Submitted to the Department of Electrical Engineering and Computer Science
on January 28, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science

Abstract

Understanding how variation in genome sequence leads to differences in gene regulation is a longstanding challenge that is essential to explaining the many phenotypic differences and complex diseases that are observed in humans. Sequencing-based functional genomics assays provide unique insight into this problem by allowing direct observation of differences between homologous chromosomes in, for example, gene expression, transcription factor binding, or chromatin state.

In this thesis, we use data from the ENCODE project to conduct a unique examination of allele-specific activity jointly across many layers of regulation including chromatin structure and modifications, occupancy by transcription factors and RNA Polymerase II, and ultimately gene expression. We develop new computational approaches for (1) creating personal genomes; (2) facilitating their use in the analysis of sequenced reads; (3) detecting allele-specific activity; (4) identifying allelic differences in transcription factor binding motifs; and (5) jointly analyzing functional data to identify putative causal variants in eQTLs or GWAS loci. We show that these approaches improve upon existing methods.

We observe that there are genome-wide correlations in allele-specific activity, and that allele-specific activity is widespread across the autosomes. We demonstrate that we can gain insights into gene regulation by combining the signals of allele-specific activity from multiple assays. By detecting variants that alter transcription factor binding we find that we can identify putative causal variants in eQTLs. We show that allele-specific activity is enriched at GWAS SNPs and eQTLs and propose how analysis of allele-specific activity in individuals could provide an alternate pathway to discovery of eQTLs or identification of causal variants in eQTLs or GWAS loci.

Thesis Supervisor: Manolis Kellis
Title: Professor of Computer Science

Acknowledgments

MIT is a community of incredible people, and I've had the good fortune to have interacted with many wonderful people along this journey. With many of these people I've forged friendships that will stand the test of time. There are too many people to name individually, but among them are fellow students with whom I've participated in extra-curricular activities, and classmates with whom I spent long hours studying and working on problem sets. I've had the privilege of being surrounded by fellow lab members who are brilliant, funny, caring, and selfless people whose insights, suggestions, and encouragement have been invaluable. In the last few years I've had the pleasure of sharing an office with Pouya, Dave, Stefan, Luke, Abhishek, Xinchun, Richard, Kunal, and Angela, and our adorable office puppy, Atlas. With my labmates and officemates I've enjoyed countless hours spent discussing science and a myriad of other topics. I've benefited from the support of a great number of friends from outside the MIT community, as well.

I'm thankful for the assistance of numerous administrators, administrative staff, and technical staff in CSAIL and the EECS graduate office, especially Bryt Bradley, Janet Fischer, and Terry Orlando, who have helped solve so many problems large and small.

I thank Pardis Sabeti and Pete Szolovits for graciously agreeing to serve on my committee and for their advice and guidance.

I am forever grateful and indebted to Manolis Kellis for giving me the opportunity to be a member of his lab and for his unwavering support. His enthusiasm is contagious and I'm continuously amazed by his energy. In addition to his scientific advice he has helped me to learn many life lessons.

Finally, I cannot thank my family enough for their endless encouragement, support, and love. My parents, Ruth and Ed, have been continuously optimistic, and helpful on many levels. My wife Jen, and son, Jacob, and our pets, have tolerated me not spending nearly as much time with them as they deserve, often seeing me for only a few precious moments in the mornings before we start our days. Nonetheless, the time we spend together, especially our walks with our dog, Remy, brings me the greatest joy.

Contents

| | | |
|----------|--|-----------|
| 1 | Background | 11 |
| 1.1 | Relevant molecular biology | 11 |
| 1.1.1 | DNA, RNA, and Protein | 11 |
| 1.1.2 | Gene expression and regulation | 14 |
| 1.1.3 | Transcription factor binding motifs | 16 |
| 1.2 | Experimental techniques | 17 |
| 1.2.1 | DNA Sequencing | 17 |
| 1.2.2 | Gene expression analysis by RNA sequencing | 19 |
| 1.2.3 | Detecting protein-DNA interactions with chromatin immunoprecipitation followed by sequencing | 20 |
| 1.3 | Computational and analytical methods | 22 |
| 1.3.1 | Sequenced Read Alignment | 22 |
| 1.3.2 | Detecting and phasing genetic variants | 23 |
| 1.3.3 | Genome Wide Association Studies and Quantitative Trait Loci | 25 |
| 1.3.4 | Hidden Markov models | 26 |
| 1.4 | Thesis overview | 29 |
| 2 | Constructing Personal Genomes | 31 |
| 2.1 | Introduction | 31 |
| 2.2 | Aligning sequenced reads to personal genomes to avoid reference bias | 32 |
| 2.3 | Haplotype assignment and creation of personal genomes | 38 |
| 2.3.1 | The simplest case: non-overlapping variants | 38 |

| | | |
|----------|---|-----------|
| 2.3.2 | Challenges of haplotype assignment | 40 |
| 2.3.3 | Maximum-likelihood haplotype assignment using a context-sensitive input/output hidden Markov model | 43 |
| 2.3.4 | Personal genome creation | 50 |
| 2.4 | Comparison of personal genomes created with PEGASUS and AlleleSeq | 51 |
| 3 | Methods for analyzing sequenced reads with personal genomes and for detecting allele-specific activity | 53 |
| 3.1 | Introduction | 53 |
| 3.2 | Incorporating a personal genome into standard workflows | 55 |
| 3.3 | Variant-aware detection of PCR duplicates | 56 |
| 3.4 | Detecting allele-specific activity at heterozygous variants | 57 |
| 3.5 | Comparison of PEGASUS and AlleleSeq for detecting allele-specific activity | 60 |
| 3.6 | Detecting allele-specific activity at functional elements | 61 |
| 4 | A genome-wide survey of allele-specific activity in a human genome | 65 |
| 4.1 | Introduction | 65 |
| 4.2 | Methods | 66 |
| 4.3 | Validation of method for detecting allele-specific activity | 67 |
| 4.4 | Genome-wide allelic correlations | 68 |
| 4.5 | Allele-specific activity is widespread across the GM12878 genome | 69 |
| 4.6 | Gaining insights into gene regulation | 71 |
| 5 | Identifying sequence variants that have functional effects | 75 |
| 5.1 | Introduction | 75 |
| 5.2 | Detecting allelic motifs | 76 |
| 5.3 | Allele-specific activity correlates with change in motif PWM score | 78 |
| 5.4 | Enrichment for allele-specific activity at GWAS loci and eQTLs | 79 |
| 5.5 | Discovering mechanisms for disease association and eQTLs | 81 |

| | |
|----------------------------------|-----------|
| 6 Conclusion | 87 |
| 6.1 Summary of results | 87 |
| 6.2 Future work | 88 |

Chapter 1

Background

1.1 Relevant molecular biology

1.1.1 DNA, RNA, and Protein

All cells rely primarily on three types of molecules to control how they function: deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and protein. The instructions for life are encoded in DNA by all known organisms. Many of these instructions specify the sequences of amino acids that comprise proteins, which perform a myriad of biological functions. RNA has several ancient and essential functions in the production of protein that have been known for decades and more recently it has been discovered to have important regulatory functions.

DNA is a polymer composed of four types of nucleotides, adenine (A), cytosine (C), guanine (G), and thymine (T). The nucleotides, also called bases, join together by their deoxyribose sugars and phosphate groups to form long, directional strands. The orientation of the strands is indicated by referencing the 5' (five prime) and 3' (three prime) carbon atoms of the deoxyribose sugars. These bases also pair with each other via hydrogen bonding, A with T, and C with G, which allows two strands to align anti-parallel to each other and form the famous double helix structure of DNA. In eukaryotic cells the resulting double-stranded DNA macromolecules that make up the organism's genome are organized into

chromosomes located in the nucleus of each cell. In order for the chromosomes, which can be hundreds of millions of base pairs (bp) long, to be compact enough to fit inside the cell's nucleus the DNA is wrapped around octamers of histone proteins, forming nucleosomes. This compact form of DNA is known as chromatin and the level of compactness may be altered by a variety of chemical modifications of the histone proteins (Kouzarides, 2007).

RNA is a polymer composed of nucleotides, like DNA, but most notably uses uracil (U) in place of thymine. It typically maintains a single-stranded form, and can also fold on itself resulting in a variety of secondary structures. Among the numerous functions of RNA within a cell, this thesis is primarily concerned with the role of messenger RNA (mRNA), which serves as a template for producing protein. The algorithms and methods described in this thesis are, however, also applicable to several classes of non-coding RNA (ncRNA) with regulatory functions.

Two of the types of functional elements in which the instructions in DNA are encoded are genes and transcriptional regulatory regions. A gene is a region of DNA specifying either the sequence of a functional ncRNA or encoding the amino acid sequence for a protein. The cell produces protein by first transcribing the region into mRNA with an RNA Polymerase protein, and then translating the mRNA into the sequence of amino acids constituting the protein. Transcriptional regulatory regions are regions of DNA that control the conditions under which genes are transcribed (also called "expressed"), and the amount of transcription that occurs.

Twenty standard amino acids can be linked together into polypeptide chains in a myriad of combinations to produce proteins. The chemical properties of the amino acids determine the three-dimensional structure of the resulting proteins and, correspondingly, their functions. The two types of proteins that are most relevant to this thesis are transcription factors (TFs) and histones. Transcription factors (TFs) are proteins that bind to the regulatory regions of DNA in order to regulate gene expression. Individual TFs may be promoters, which activate

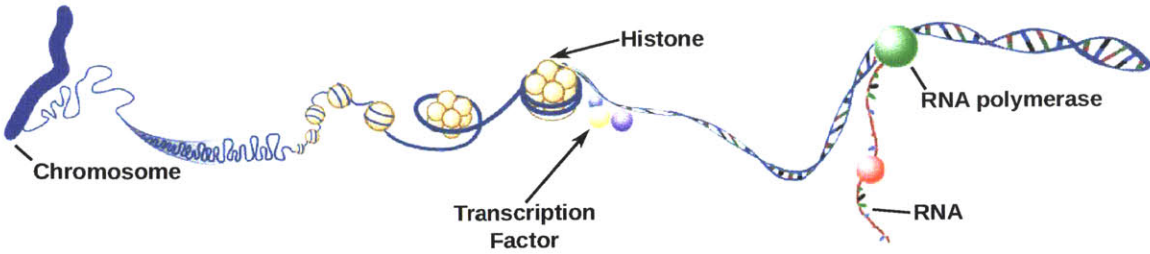


Figure 1-1: Chromosomes are composed of DNA that has been wrapped tightly around histone octamers and organized into a highly compacted form. At regions where the chromatin has a more open configuration the DNA is accessible and can be bound by transcription factors. Transcription factors can recruit RNA Polymerase to transcribe genes into RNA, which may have its own function or be translated into protein. (Image adapted from ENCODE)

or increase expression, or repressors, which inhibit or decrease expression, or may take on both roles depending on the context. Histone proteins form octamers which DNA wraps around, as mentioned previously, and their individual amino acids are frequently observed to have chemical modifications including acetylation, methylation, and phosphorylation. These modifications can alter the compactness of the chromatin and the accessibility of the DNA, and the modified histones may also interact directly with other proteins, for example to regulation gene expression. Dozens of types of these modifications have been observed, and understanding their functions is an active area of research. A simple example of the structure and interaction of these proteins with DNA and RNA is shown in figure 1-1.

In the case of all three of these molecules, DNA, RNA, and protein, seemingly minor chemical changes can cause profound differences in function. In the case of protein, a substitution of a single amino acid may in some situations increase the activity of the protein, and in others result in complete lack of function of the protein. In certain instances these amino acid substitutions may result from a change as minimal as a single nucleotide substitution in the underlying gene, or in the transcribed RNA. Also, in the case of transcriptional regulatory regions, such a minimal change in the DNA can substantially alter the chemical affinity

with which a transcription factor binds, resulting in significant changes in the frequency of gene transcription (Seto et al., 1991; Matys et al., 2003; Kasowski et al., 2010).

1.1.2 Gene expression and regulation

While an organism has a single genome which is shared by all of its cells, the selection of genes that are expressed determines the function of a cell and distinguishes the hundreds of different types of cells from each other. Understanding the process by which genes are selected to be expressed in particular cells, and specifically how variation in genome sequence leads to differences in gene regulation is a longstanding challenge that is essential to explaining the many phenotypic differences and complex diseases that are observed in humans.

In order to produce a functional protein or ncRNA, a gene must first be transcribed into RNA in its entirety by an RNA polymerase protein. This also requires the chromatin to be in an "open" configuration that makes the region of the gene accessible to RNA polymerase and other transcriptional machinery. The general structure of protein-coding genes in eukaryotes is shown in figure 1-2. The nucleotide positions where RNA polymerase starts and finishes transcribing the gene are known as the transcription start site (TSS) and transcription end site (TES), respectively. RNA is always synthesized in the direction of 5' (five prime) to 3' (three prime). Accordingly, transcription always occurs in the same direction relative to the strand of DNA being transcribed; RNA polymerase proceeds along the template DNA strand in the 3' to 5' direction assembling a complementary RNA strand in the 5' to 3' direction. The transcribed region of the gene, also called the gene body, begins with a 5' untranslated region (5' UTR), and ends with a 3' untranslated region (3' UTR), and in between is typically composed of expressed regions (exons) encoding the amino acid sequence of the protein separated by non-coding intragenic regions (introns), which are spliced out of the RNA before it is translated to produce a protein (genes can consist of just a sin-

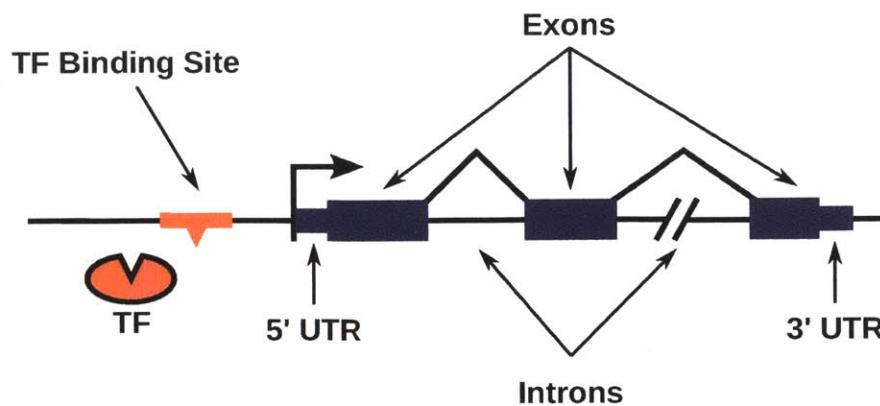


Figure 1-2: A simple model of the structure of a gene. The transcription start site and direction of transcription are indicated by the arrow pointing right located immediately before the 5' UTR (short blue rectangle). Exons are shown as larger blue rectangles, separated by introns. The 3' UTR marks the end of the transcribed portion of the gene. A transcription factor binding site located in the promoter of the gene, and a TF are also depicted.

gle exon, however). Within the exons each three consecutive nucleotides, called a codon, specify one amino acid. Like protein-coding genes, ncRNA genes also have a TSS and TES. Some long non-coding RNA genes have even been shown to have multiple exons and introns (Guttman et al., 2010). In ncRNAs the exons obviously do not encode amino acids, but the introns are still spliced out to produce the mature, functional ncRNA.

There are numerous stages at which gene expression may be regulated, and this thesis will focus on pre-transcriptional regulation. The region of DNA that is of central importance to gene regulation is the gene's promoter, which immediately precedes the TSS. The promoter contains sequences that are bound by TFs serving to activate or repress transcription of the gene by RNA polymerase. TFs may also regulate genes by binding at sites within the gene body or at enhancer regions, which are known to be as far as a million bases away from the TSS (Figure 1-3). Activators most commonly function by interacting directly with RNA polymerase, or with other proteins that are involved in the binding of RNA polymerase and initiation of transcription, and facilitating those processes. Activators

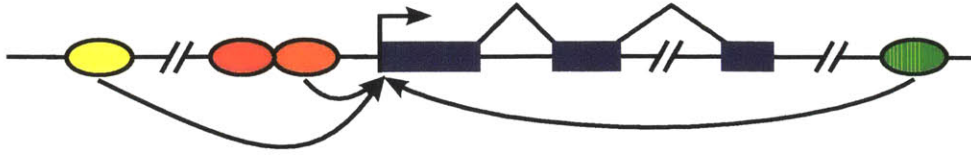


Figure 1-3: A simple structure representing a gene is depicted as in Figure 1-2. Although that example showed a single TF binding site at the promoter, in actuality, genes are regulated by numerous transcription factors that can bind upstream and downstream of the TSS, and in introns, and that may bind millions of bases away from the TSS.

may also function by opening up the chromatin and making it more accessible to other TFs and RNA Polymerase. Repressors, on the contrary, decrease the frequency of transcription by blocking RNA Polymerase or activators from binding to the DNA.

1.1.3 Transcription factor binding motifs

Many transcription factors have a protein structure such that the amino acids in the DNA-binding domain chemically recognize particular, short sequences of DNA bases. The transcription factors then bind selectively to those particular DNA sequences, which are enriched in the regulatory regions of each of the genes they regulate. The patterns of DNA bases that describe these recurring, short sequences, often 5-15bp long, are known as transcription factor binding motifs, and are often represented with logos (Figure 1-4).

The regulatory functions and properties of transcription factors and methods for detecting binding motifs have been studied extensively (Kheradpour and Kelis, 2014). The property of TF binding motifs that is most relevant to this thesis is the relationship between DNA sequence and binding affinity of the TF. It is commonly the case that positions in the motif sequence show a high specificity for just one or two specific DNA bases. Accordingly, the alteration of even a single DNA base in a TF binding site can interfere with the ability of the TF to bind at that site. That, in turn, can lead to a cascade of effects resulting in a change in the expression of the gene being regulated by the affected TF. Accordingly, when

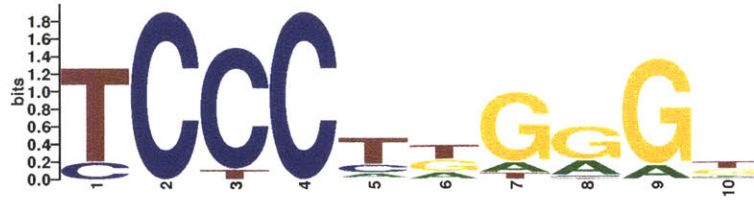


Figure 1-4: A logo for a motif for the transcription factor EBF1. The DNA bases that are observed in instances of the motif sequence are shown stacked on top of each other at each position. The height of the stack indicates the information content of that position of the motif sequence, while the height of each base indicates the frequency with which it's observed at that position. For example, "T" is the base most frequently observed at the first position, but at some locations where the TF binds a "C" is observed instead. "C" is the only base observed at the second and fourth positions of the motif sequence. The fifth and sixth positions of the motif sequence show low specificity and several different bases are observed at these positions.

trying to understand the function of variants in non-coding regions it is helpful to be able to detect both when those variants alter instances of TF binding motifs, and whether the binding of the TF is altered.

1.2 Experimental techniques

1.2.1 DNA Sequencing

The development of technology to efficiently determine the sequence of bases in a DNA molecule launched the ongoing genomics revolution. Advances and automation of Sanger sequencing, which was developed in the 1970s (Sanger et al., 1977), enabled the completion of the Human Genome Project in 2003. Sanger sequencing remains the gold standard for minimizing errors, but Next Generation Sequencing (NGS) methods that employed a variety of other technologies became available starting in the 2000s and have both become the standard for whole genome sequencing and enabled a wide range of sequencing based assays. As NGS methods have continued to improve it has become possible to sequence entire genomes in a single day and for a cost of under \$1000, several

orders of magnitude faster and less expensive than Sanger sequencing (Dondorp and de Wert, 2013; van Dijk et al., 2014).

Presently, neither Sanger sequencing, nor NGS methods, nor methods based on even newer technologies are capable of sequencing a DNA molecule as long as a human chromosome in its entirety as a single sequenced "read". Rather, for differing technological reasons all of the methods are limited to sequencing relatively short lengths of DNA. In the case of Sanger sequencing, with current machines, the sequenced "reads" are limited to at most 1000 bases and at most 384 DNA samples can be sequenced in parallel. Of the numerous NGS technologies, we'll focus on Illumina (Solexa) Sequencing, because it's the technology used for many of the assays analyzed in this thesis. Although Illumina Sequencing technology has improved since those assays were performed, even the current Illumina sequencing machines are limited to producing reads up to about 300 bases long. They can, however, sequence hundreds of millions of samples at once (van Dijk et al., 2014).

Although the technologies and protocols used by the Sanger and Illumina sequencing methods differ, they also have a number of features in common. For both methods the DNA molecules to be sequenced must be broken into suitably sized fragments. The lengths of the fragments can range from several hundred bases to thousands of bases, and typically fragments of one particular length are selected for sequencing. Both methods involve synthesizing a DNA strand that is complementary to one strand of the input DNA molecule for which the sequence is desired. The complementary strand of DNA is synthesized in the 5' to 3' direction beginning at the 3' end of one strand of the input molecule. The use of differently colored fluorescently labeled nucleotides allow for the sequence of the synthesized strand to be observed and reported as a "single-end" read. Modifications of the protocols make it possible for the sequence of the DNA molecules to be read from both ends yielding paired-end reads.

For the purpose of whole genome sequencing these short reads must be assembled to reconstruct the continuous sequences of the chromosomes. This can

be done using only the sequenced reads, a process called *de novo* assembly. Alternatively, when a complete genome sequence is already available for the species being sequenced (or a closely related species), it may be used as a reference for a process called template-based assembly. In both cases paired-end reads are particularly useful, because the combination of the sequences and an approximate length of the fragment they came from is helpful for resolving ambiguities about the arrangement of the reads in the continuous sequence.

The ability to determine the reference genome sequences of species and the genome sequences of individuals has led to the use of computational methods for identifying functional elements (Consortium et al., 2012b) and performing detailed comparisons between species (Lindblad-Toh et al., 2011) and across human populations, including detecting sequence variation and measuring the frequency of variations across populations (Consortium et al., 2012a). Furthermore, modern DNA sequencing technology has enabled a variety of sequencing-based assays for measuring the functions of cells.

1.2.2 Gene expression analysis by RNA sequencing

Ideally, the selection of genes that are expressed by a cell would be determined by directly measuring the quantities of all of the proteins present in a cell, but technical limitations make this extremely difficult (Garbis et al., 2005). Fortunately, although gene expression is also regulated by post-transcriptional processes, the sequencing of transcripts can be used as an indicator of gene expression in place of direct measurements of proteins. As well, sequencing of RNA molecules has benefited from all of the advancements in DNA sequencing.

A variety of methods exist for extracting RNA from one or more cells, and allow for selecting RNA molecules of a particular length, a particular type such as mRNA, or from a particular compartment of the cell (Rosenbloom et al., 2010). A reverse transcriptase enzyme can then be used to synthesize complementary DNA molecules (cDNA) for the extracted RNA. The sequence of the RNA can

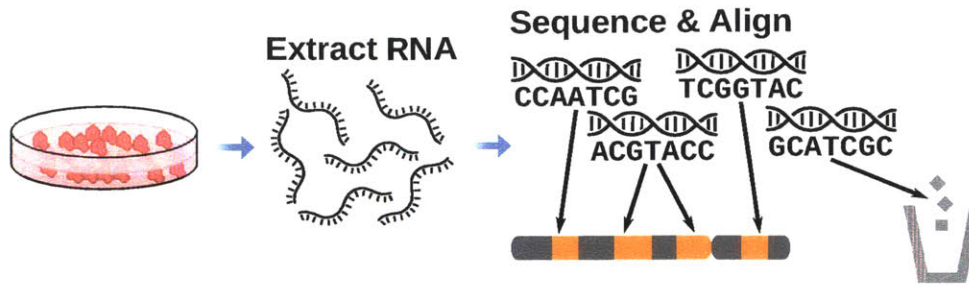


Figure 1-5: An RNA-Seq assay is performed by extracting RNA from a sample (or culture) of cells. A reverse transcriptase protein is used to transcribe the RNA to complementary DNA (cDNA), which is then sequenced. The sequenced reads are aligned to the genome. The genes that the cells are expressing are indicated by the locations where the reads align.

then be determined by sequencing the cDNA using any of the available methods for DNA sequencing. Although the short reads produced by this process of RNA Sequencing (RNA-Seq) typically will not be transcripts of complete genes, when the genome sequence of the species (or individual) is known the reads can be aligned back to the genome to reveal the genes from which they were transcribed (Figure 1-5) (Wang et al., 2009). Statistical analysis of the reads can reveal not only whether a gene is expressed, but also the relative levels of expression of different genes (Trapnell et al., 2012). Furthermore, for diploid cells, that contain two copies of each chromosome, and for genes which contain variants in coding regions it's possible to determine whether a gene is being expressed in similar quantities from each chromosome, or preferentially from just one chromosome. The latter situation is known as allele-specific expression (ASE) (Degner et al., 2009).

1.2.3 Detecting protein-DNA interactions with chromatin immunoprecipitation followed by sequencing

In addition to detecting the genes that are expressed in cells of a particular type, it is important to understand why they are expressed. Determining the epigenetic histone modifications and sets of transcription factors that are involved in the

regulation of genes and detecting where and when they are acting is critically important to understanding gene regulation. Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) is a method that makes it possible to identify locations throughout the genome where a target protein, such as a transcription factor, or a histone with a particular chemical modification, is binding to the DNA (Jothi et al., 2008; Gossett and Lieb, 2010). In a ChIP-Seq assay the cells are first typically treated in a manner that causes reversible cross-links to form between the DNA and the proteins that are bound to the DNA, for example using formaldehyde. The DNA is then sheared to produce fragments with an average size of about 500bp. An immunoprecipitation step is then performed using an antibody specific to the target protein to enrich for fragments to which the target protein is cross-linked. After reversing the cross-links the fragments of DNA can be sequenced. Then, as in the case of RNA-Seq, the sequenced reads can be aligned back to the genome to determine where the target protein was bound.

Although it is advantageous to have paired-end reads or relatively long reads for whole genome sequencing or RNA-Seq, for ChIP-Seq shorter, single-end reads are typically sufficient for identifying a unique location in the genome where a read aligns with high confidence. When using Illumina sequencing, the length of the sequenced reads that are produced can be controlled because it corresponds directly to the number of times the chemical reactions used for sequencing are performed. Many of the reads in the data analyzed in this thesis are 36 bases long and were sequenced using Illumina machines, for example.

The immunoprecipitation process produces a sample of DNA fragments that are enriched for fragments cross-linked to the target protein, but the sample will contain many extraneous fragments of DNA as well, because of the nature of the process. It is, therefore, necessary to use a peak calling algorithm to distinguish the locations where the target protein is most likely to be bound from locations to which reads aligned as a result of experimental noise inherent to the assay (Figure 1-6) (Guo et al., 2010). As in the case of RNA-Seq, when reads overlap regions of the genome containing a variant it is possible to determine whether or not the

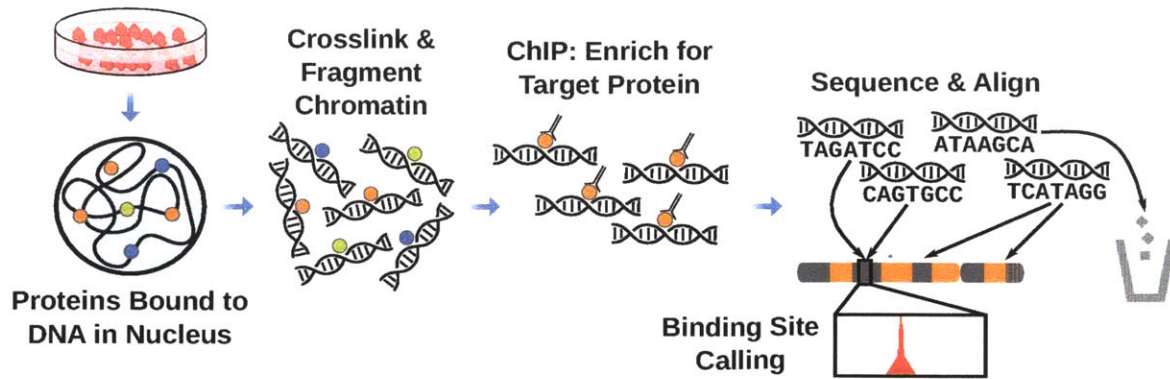


Figure 1-6: A ChIP-Seq assay requires a number of additional steps as compared to an RNA-Seq assay. Cells are treated to create chemical cross-links between DNA and the proteins that are bound to it. The chromatin is extracted and sheared into small fragments. A chromatin immunoprecipitation step is performed to enrich for fragments with a target protein. Aligning the sequenced reads to the genome reveals the locations where the target protein was bound, but some fragments that were not bound by the protein will also end up getting sequenced. A peak calling algorithm is used to detect binding sites amidst the background noise resulting from the unbound fragments.

protein is binding similarly frequently to both chromosomes or preferentially to one chromosome, known as allele-specific binding (ASB).

1.3 Computational and analytical methods

1.3.1 Sequenced Read Alignment

A critically important, and challenging, step in the analysis of reads from sequencing based assays, is the alignment of the reads to the genome. A typical assay might produce tens or even hundreds of millions of reads on the order of 30-100bp long, and aligning the reads requires searching for each read sequence within a much longer genome sequence (the human genome is about 3 billion base pairs long). The software programs that have been developed to perform this task efficiently typically fall into two categories: methods based on the Burrows-Wheeler transform (Langmead et al., 2009; Li and Durbin, 2009), and hashing-based methods (Li et al., 2008b,a). In both cases the most challenging aspect of

the task is accurately aligning reads that do not perfectly match the sequence of the genome. Aligning reads which don't perfectly match the genome sequence requires a significant amount of time, and as more mismatches are allowed it becomes more likely that a read will align equally well to multiple locations in the genome, which limits its informative value.

One reason that a read sequence may not match the genome sequence is because bases can be called incorrectly when the DNA is being sequenced. The different sequencing technologies have different error modes, and these errors are, in fact, relatively common; sequencing machines typically output a quality score for each base, and an average quality score corresponding to a 1

1.3.2 Detecting and phasing genetic variants

The human genome is about 3 billion bp long, of which about 1.5

Common types of genetic variation include small-scale variations such as single nucleotide polymorphisms (SNPs), and insertions and deletions ("indels", collectively), and larger structural variations, typically over 1000bp, such as long deletions, copy number variants, and translocations (Figure 1-7). This thesis is primarily concerned with the analysis of SNPs and indels, which are relatively more common than the other types of variants, and which may have effects that are less easily explained.

Although microarray chips, for example, are a reliable and efficient tool to detect known variants, especially SNPs, in samples of DNA, detecting all the variants in an individual's genome is challenging, especially for types of variants other than SNPs. Numerous methods have been developed to detect variants using sequenced DNA reads (Nielsen et al., 2011). The particular explanation of sequence variations reported by any given variant calling method will depend on both the design of the algorithm as well as the specific parameters that are selected. As well, some variant calling methods are designed for detection of a single specific type of variant, for example indels, and so will only use that

| | |
|--------------------------------|--|
| Single Nucleotide Polymorphism | AGTGTCCG T GCTGTGGG AGTGTCCG G GCTGTGGG |
| Insertion-Deletion | AGTGTCC GTG CTGTGGG AGTGT C -----CTGTGGG |
| Inversion | AGTGTCC GTG CTGTGGG AGTGTCC CACTG GTGGG |
| Copy Number Variant | AGTGTCC GTGCCGTG CTGTGGG AGTGTCC GTG -----CTGTGGG |

Figure 1-7: Examples of several types of common genetic variants: A single nucleotide polymorphism (SNP) alters a single DNA base. An insertion-deletion (“indel”) adds or removes one or more DNA bases. An inversion substitutes a sequence of bases with its reverse complement. A copy number variant (CNV) occurs when there can be a variable number of copies of a particular sequence.

type of variant to explain observed sequence variation. Accordingly, when trying to detect variants genome-wide it is common to use multiple variant calling technologies and methods, and for those methods to produce variant calls that overlap and may be inconsistent with each other

In addition to detecting variants, for diploid organisms there is an additional challenge of "phasing" the variants, the process of determining which allele of each variant belongs on each of the two homologous chromosomes. When the variant calls are known for an individual organism and the parents of the individual then this problem can mostly be solved by trio-phasing (Figure 1-8). Trio-phasing is ideal because it results in an assignment of each copy of a chromosome to a parent and a consistent assignment of variant alleles across the entire chromosome. Often the genome sequences of the parents are unavailable, but if sequenced DNA reads are available, either from whole genome sequencing or sequencing based assays, then a technique known as read-backed phasing may be used. Read-backed phasing algorithms work by identifying sequenced reads that overlap multiple variants. The alleles occurring within the read can then be assigned to a phase set that indicates they are part of a single haplotype and occur on the same chromosome as each other (Figure 1-9). In the same way that contigs are constructed during de novo assembly, multiple reads overlapping the same variant alleles may be linked together to expand the phase set. Read-

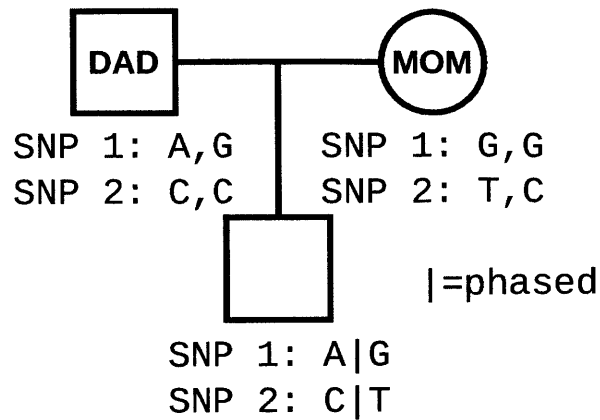


Figure 1-8: A pedigree diagram indicates the relationship of a child and his parents, along with the genotype of each individual for two SNPs. At “SNP 1” the father is heterozygous with an “A” allele and a “G” allele, and the mother is homozygous for the “G” allele. Their son is also heterozygous with the “A” and “G” alleles. The child must have inherited the “A” allele from the father and the “G” allele from the mother, because the mother does not have an “A” allele. Similarly, for “SNP 2”, the child must have inherited the “T” from his mother and the “C” from his father.

backed phasing is most effective with paired-end reads, but may be performed with single-end reads, too.

1.3.3 Genome Wide Association Studies and Quantitative Trait Loci

Genome Wide Associate Studies are a powerful approach to identifying genetic variants that are associated with a specific phenotype, such as a disease (Altshuler et al., 2008; Hirschhorn and Daly, 2005). In recent years GWAS studies have identified hundreds of variants associated with height (Wood et al., 2014) and hundreds more associated with a variety of diseases (of the Psychiatric Genomics Consortium et al., 2014; Lambert et al., 2013; Morris et al., 2012; Schunkert et al., 2011). By their nature GWAS studies identify variants that are not necessarily causal for the phenotype, but that merely serve as markers for regions within the genome that are associated with the occurrence of the phenotype. Additional research must then be conducted to discover a mechanism by which a particular


```

Read 1: AGTGTCC-----TGCTGTAG
Haplotype 1: AGTGTCCCTCAAACCTCGGTAACGCATATACCGCTGTAGG
Reference: AGTGTCCCTCAAACCTCGGTAACGCATATACCGCTGTAGGG
Haplotype 2: AGTTTCCCTCAAACCTCGGTAACGCATATACCGCTGTGGG
Read 2:  TTTCCCTC-----GCTGTGGG
Phase Set:
G|T,A|G

```

Figure 1-9: A section of DNA sequence from each of two homologous chromosomes is shown above and below the reference sequence. The locations of two SNPs are indicated by red bars. The sequence of “Read 1” contains a “G” at the location of the leftmost SNP, and an “A” at the other SNP. These alleles must occur on the same haplotype (“Haplotype 1”), because the read was produced from a single molecule of DNA. Similarly, “Read 2” contains a “T” allele for the leftmost SNP, and a “G” allele for the other SNP, so those alleles must be on the same haplotype. If a third read contained a “G” at the position of the first SNP, and a “C” at the position of a third SNP (not shown), then the “C” allele would be constrained to be on “Haplotype 1”, too.

GWAS locus may affect a phenotype. This is a challenging and time consuming process because GWAS loci may be more than a million bases long and contain dozens of genes and countless regulatory elements.

A closely related approach called an expression quantitative trait loci (eQTL) study can be used to identify variants that are associated with differences in expression of genes (Rockman and Kruglyak, 2006). eQTL studies can provide valuable information about the regulatory control of a gene and the regions involved, but like GWAS studies the variants identified from an eQTL study serve only as markers for the associated regions, and additional analysis is needed to identify any causal variants and mechanisms of action.

1.3.4 Hidden Markov models

Hidden Markov Models (HMMs) are a type of computational model that are frequently used for analysis of biological sequences and genomic data (Durbin et al., 1998) among many other uses. Formally, for a set of observed outputs $y = \{y_1, \dots, y_N\}$ from an alphabet $V = \{v_1, \dots, v_{|V|}\}$, an HMM is defined by a sequence of states, $Q = \{q_1, \dots, q_N\}$ drawn from a state alphabet $S = \{s_1, \dots, s_{|S|}\}$, a matrix of transition probabilities between the states, $A \in \mathbb{R}^{(|S|+1) \times (|S|+1)}$, and an emission matrix indicating the probability of emitting each element of V

from each state s , $B \in \mathbb{R}^{(|S|+1) \times |V|}$. A graphical model diagram for a standard HMM is shown in Figure 1-10a. A very useful function of an HMM is that when only the observed sequence of output is known, and the state the system was actually in at each position (or timepoint) of the sequence is unknown, the Viterbi decoding may be used to determine a maximum likelihood state sequence: $\operatorname{argmax}_{\vec{q}} P(\vec{q}, \vec{y}; A, B)$. The Viterbi decoding can be calculated efficiently by making use of the independence property and using dynamic programming to implement the algorithm.

A central feature of a standard HMM is that it is memoryless, with the future state of the system depending only on the current state, and not on the state of the system at any previous position. Although memory may effectively be included in an HMM by the addition of states to the model, this can quickly become unwieldy when modeling a complicated system. Instead, this addition of memory to an HMM has been formalized by the definition of a context-sensitive HMM (Yoon and Vaidyanathan, 2004), which describes how a limited amount of memory may be associated with the states of the HMM (Figure 1-10b).

Another feature of a standard HMM is that it takes no input and emits just a single observation at each position. These abilities, too, have been formalized in the definition of an input/output HMM (IOHMM) (Bengio and Frasconi, 1995), which, as the name implies, can read input at each state, and produce an output according to both that input and the current state of the model (Figure 1-10c).

The context-sensitive HMM and the IOHMM may be combined together resulting in an HMM-variant that in addition to accepting input and can generate output based on that input, the included memory, and the current state of the model. The context-sensitive IOHMM and a modified version of the dynamic programming algorithm for computing the Viterbi decoding are described in detail in Chapter 2.

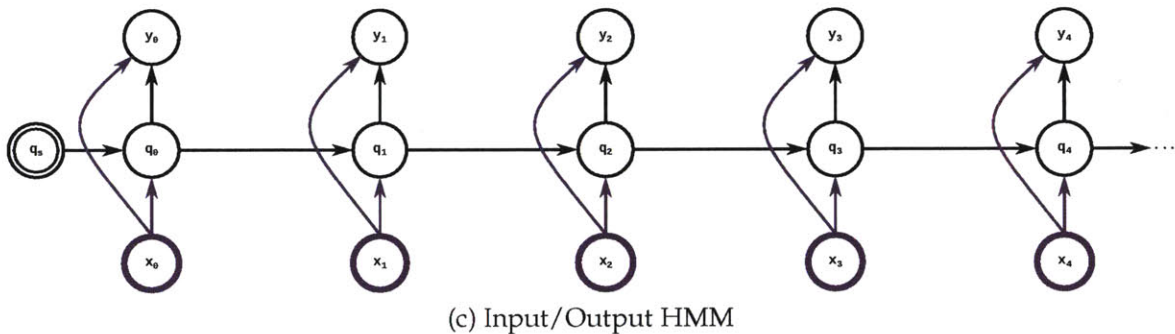
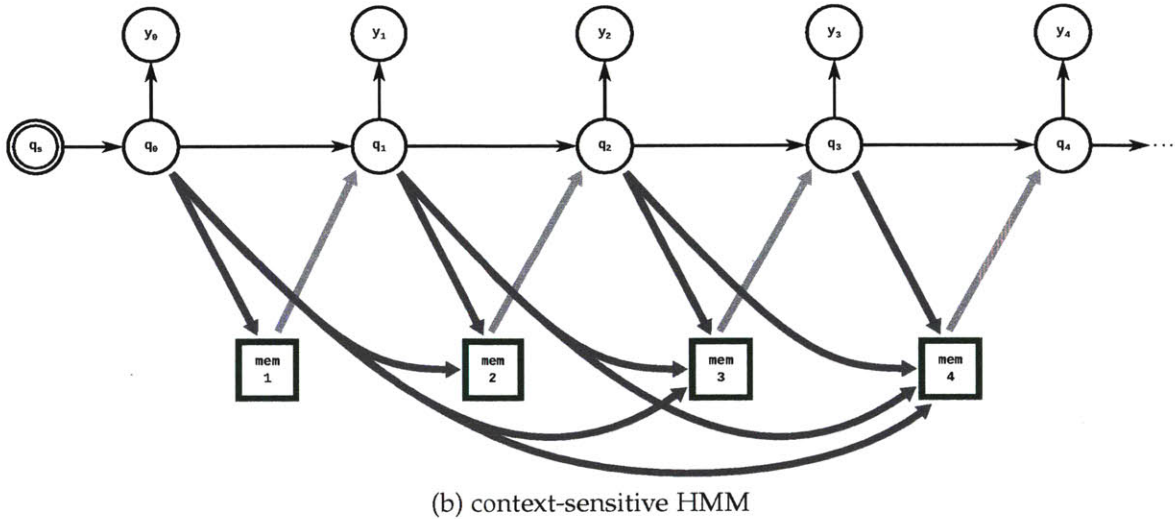
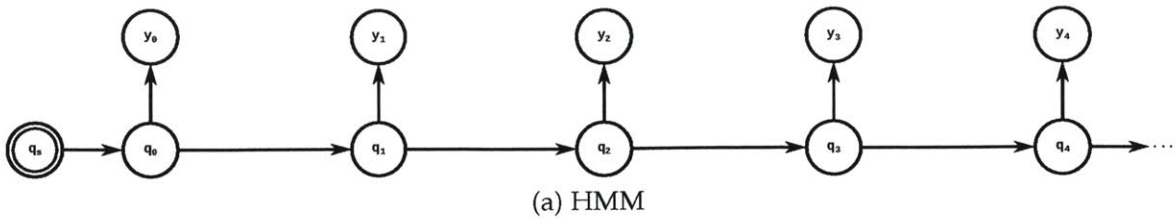


Figure 1-10: (1-10a) The graphical model diagram for a standard HMM showing a sequence of states “ q_i ” and the corresponding observations “ y_i ”. (1-10b) The graphical model diagram for a context-sensitive HMM adds memory (green box) associated with each position in the state sequence. Each state can write to any subsequent block of memory (dark gray arrows), and can read from the one block of memory associated with it as indicated by the light gray arrow. (1-10c) An Input/Output HMM adds an input “ x_i ”, capable of influencing the output, to each state in the sequence.

1.4 Thesis overview

This thesis describes computational approaches for detecting and understanding the functional effects of sequence variants, and the results obtained by analyzing sequencing based functional assays using a personal, diploid genome. The contributions include:

- A method for constructing personal, diploid genomes, that incorporate the variants that have been detected in an individual's genome sequence, which is the central component of the PEGASUS software package (Chapter 2). This task is accomplished by using a context-sensitive IOHMM, which is a novel variation of an HMM constructed by combining a context-sensitive HMM and an IOHMM. This model allows for the inconsistencies in variant calls to be efficiently resolved in a manner that maximizes the likelihood of diploid genome sequence being correct. Modifications required for the Viterbi decoding algorithm are also described.
- Methods for analyzing data from sequencing based functional assays using a diploid genome, and detecting allele-specific activity in those data, implementations of which are also included in PEGASUS (Chapter 3). These include utilities for: variant-aware PCR duplicate marking, counting allelic reads, aggregating allelic reads across regions using phasing information, detecting allelic motif differences, and working with genomic regions in multiple coordinate systems.
- The results of a genome-wide analysis of allele-specific activity in nearly 200 ENCODE datasets for the GM12878 cell line (Chapter 4). We show that allele-specific activity is widespread throughout the genome, and that there are genome-wide correlations in allelic activity among RNA Polymerase II, and dozens of histone modifications and transcription factors.
- A systematic approach for identifying sequence variants that cause allele-specific transcription factor binding and which are likely to be causal vari-

ants in GWAS loci and eQTLs (Chapter 5). This includes a method for detecting TF binding motifs that are disrupted by sequence variants using personal genomes and demonstrate that the change in PWM score is correlated with allele-specific activity. We show that GWAS loci and eQTLs are enriched for allele-specific activity. Furthermore, we show that we can detect variants associated with changes in gene expression in both proximal and distal gene regulatory regions, We demonstrate how we can use this technique to identify putative causal SNPs for eQTLs, and describe how this method could be applied to identify putative causal variants in GWAS loci as well.

Chapter 2

Constructing Personal Genomes

This chapter describes a method for creating personal, diploid genomes by producing maximum-likelihood assignments of variants to haplotypes. We describe how the use of variant calling algorithms that are specialized for particular types of variants and general uncertainty in the variant calling process results in variant calls that overlap with each other, and in some cases conflict. We present a novel variation of an HMM, the context-sensitive Input-Output HMM, along with a modified version of the Viterbi decoding that can be used to efficiently resolve overlapping and conflicting variant calls and produce a maximum-likelihood assignment of variants to haplotypes. This method is implemented as part of the Personal Genomes and Allele-Specific Utilities (PEGASUS) software package, and the personal genomes that were created with PEGASUS were essential for the analysis of allele-specific activity presented in Chapters 4 and 5. We also present the results of a comparison between the personal genome creator of PEGASUS and another tool for creating personal genomes that is part of AlleleSeq.

2.1 Introduction

Genomics assays based on short read sequencing, such as ChIP-Seq, RNA-Seq, and DNase-Seq, have become an indispensable tool for genomic analysis, useful for characterization of cellular activity, as well as comparisons over time, and

across cell types, and individuals. Furthermore, when research is conducted on important diploid organisms such as humans these assays also enable measurement of the effects of genomic variation by examining the actual alleles occurring in the sequence reads (McDaniell et al., 2010; Montgomery et al., 2010). Comparing functional data at two alleles in the same cellular environment is especially beneficial for elucidating the functional consequences of sequence variation, because it eliminates the need to control for many external and environmental variables.

The use of these assays has become widespread due to both greater availability of sequencing machines and decreases in the costs of performing the assays. Although whole genome sequencing has also become increasingly accessible and affordable, few published studies have sequenced the actual genome of the cell line or individual on which these assays are performed, or analyzed the data from the assays using that genome. Instead, a reference genome for the species is still the most common choice for alignment despite the increased accuracy that would be provided by the true diploid genome, containing all the SNPs, indels, and structural variants. Unfortunately, this results in a less accurate and less complete alignment of reads, which is a limiting factor for the quality and results of downstream analyses. For example, a recent study demonstrated that failure to correctly align and process RNA-Seq data contributed to only 12.9% of 1300 candidate loci being independently validated in a study of imprinted gene expression (Gregg et al., 2010; DeVeale et al., 2012).

2.2 Aligning sequenced reads to personal genomes to avoid reference bias

Reference genomes are haploid sequences, and although short read aligners are designed to handle small numbers of mismatches, correct reads containing SNPs with non-reference alleles often fail to align, because of the mismatches between

the true genome and the reference genome (Figure 2-1). This reduces the total percentage of reads that align at heterozygous locations and results in a bias for alignment of reads containing the reference allele (Degner et al., 2009). Furthermore, at homozygous non-reference locations a binding event could be completely missed because the signal is too weak after reads that contain variants are excluded for failing to align. Some simple methods for reducing this reference bias while still aligning to a single reference genome have been suggested, but these approaches suffer from other problems. For example, increasing the number of allowed mismatches reduces the bias, but also increases the percentage of reads that don't map uniquely (Stevenson et al., 2013), while increasing the CPU time required for alignment. Alternatively, masking the variants can cause unique regions to become non-unique, or cause reads to align better to other locations in the genome (Degner et al., 2009). Indels pose a problem for the same reasons, and the problem is further complicated by the limited ability of short read aligners to map reads with indels. Even the most capable aligners can only handle indels up to about 30bp in length (Liu et al., 2012).

As compared to using a single modified reference genome, aligning reads to a diploid genome sequence containing all the known variants permits choosing the short read aligners that work best for the types of reads being processed and maximizes the number of reads that can align (Stevenson et al., 2013) (Figure 2-3). Most importantly, aligning to a diploid genome avoids the problem of reference bias (Rozowsky et al., 2011). This is especially important for detecting allele-specific activity, because although the effect of reference bias may be small when averaged over the entire genome, at individual loci we find that it can be extremely significant (Figure).

Working with diploid genomes is a challenging problem, however, as it requires generating a diploid sequence, and management of data in three genomic coordinate systems: the two diploid haplotypes, and the reference. While some methods exist for the creation of diploid genomes (Rozowsky et al., 2011; Rivas-Astroza et al., 2011) they lack several important features. First, other software

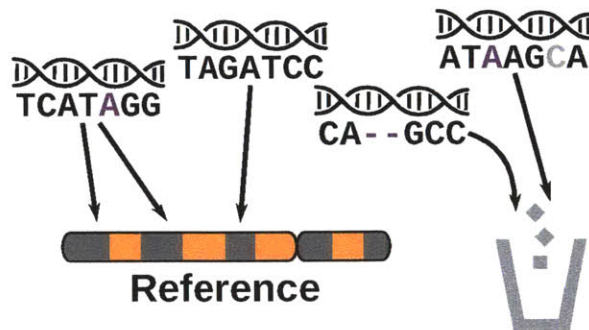


Figure 2-1: Reference bias occurs because the reference sequence is haploid and contains only the reference allele of each variant. Aligners need to treat reads with alternate alleles as having mismatches, but this causes reads with alternate alleles to be less likely to align properly than reads that are otherwise identical except for having the reference allele. The leftmost read contains the alternate allele of a SNP (shown in purple). When the aligner allows for enough error for the read to align to the correct location with a mismatch at the alternate allele, the read is also more likely to align to other locations with a mismatch at a different base. Reads containing indels (second from right) or variant alleles accompanied by sequencing errors (gray base in rightmost read), are less likely to align to the reference genome, because the mismatches required to align them are more than the amount of error that the aligner will tolerate. As a result these reads are not included in further analysis.

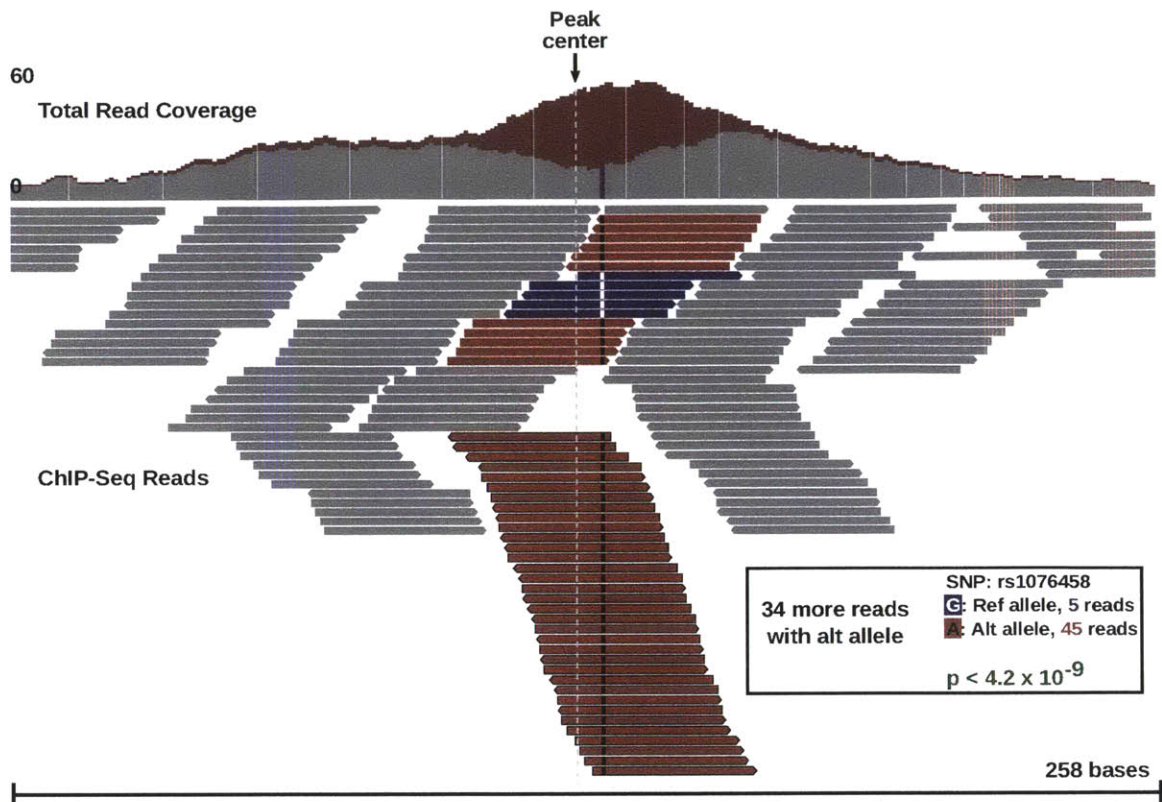


Figure 2-2: The level of read coverage at a ChIP-Seq peak for the TF EBF1 from an assay on the GM12878 cell line, with individual reads (wide rectangles) shown below the coverage signal track. A SNP alters the reference sequence just to the right of the peak center (purple bar in read coverage track). Five reads aligned to the reference sequence and containing the reference allele are highlighted in blue with the SNP position indicated in white. An additional eleven reads containing the alternate allele also align to the reference sequence (directly above and below blue reads, reads highlighted in red, SNP position in black). Aligning the same sequenced reads to the GM12878 personal genome reveals an additional thirty-four reads containing the alternate allele (bottom middle, highlighted in dark red with black outline) that failed to align to the reference sequence. The additional signal is also shown in red above the gray signal from the reads aligned to the reference. Examining only the reads that aligned to the reference sequence would lead to a conclusion that there is not a significant level of allele-specific activity ($p = 0.210$). In contrast, when all the reads aligned to the personal genome are examined it is clear that there is allele-specific binding of EBF1 to the maternal chromosome ($p < 4.2 \times 10^{-9}$).

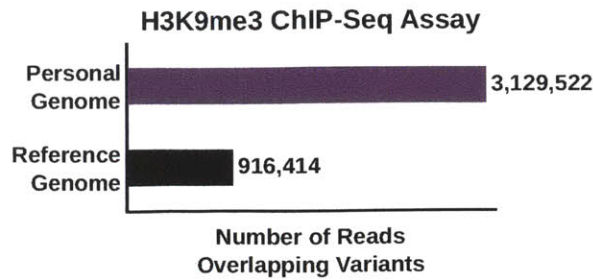


Figure 2-3: In an H3K9me3 ChIP-Seq data set from the Roadmap Epigenomics Mapping Consortium aligning reads to a personal, diploid genome results in more than three times as many reads that overlap variants aligning uniquely as compared to aligning the same reads to a reference genome.

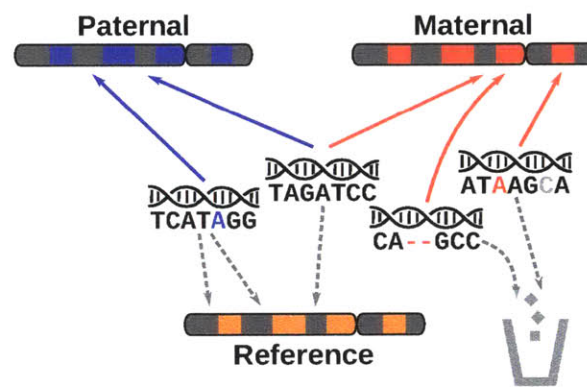


Figure 2-4: Reference bias is eliminated when reads are aligned to a personal genome because both alleles of each variant are included in the sequence. Reads that might not align to a reference sequence, as in Figure 2-1, are more likely to align when mapped to a personal, diploid genome.

offers limited control over how variants are incorporated into the sequences. For example, when the true haplotype assignments of variants are unknown the haplotypes are assigned at random, but this can result in one or more variants being left out unnecessarily when multiple variants overlap. Two more troublesome issues are that the software programs output neither the haplotypes to which heterozygous variants are assigned when the phasing is unknown, nor occurrences of overlapping SNPs and indels. All together this causes further analysis of the aligned reads to be needlessly difficult.

Aside from the two aforementioned methods for creating diploid genomes, other methods for aligning reads to modified reference genomes or detecting

allele-specific activity are computationally expensive (van de Geijn et al., 2015), come as a monolithic tool (Turro et al., 2011), performs just a single related function (Rivas-Astroza et al., 2011), require the use of a specific short read aligner (Pandey and Schlötterer, 2013), or are designed only for RNA-Seq or crosses of inbred lines (Turro et al., 2011; Pandey and Schlötterer, 2013)

To address the need for better software for working with diploid genomes, we have developed PEGASUS, a collection of software tools that facilitates creating and working with personal genomes and sequence reads aligned to them, particularly for analysis of allele-specific activity. PEGASUS is designed for flexibility, and although the components can be used in a standalone fashion, they use standardized file formats whenever possible to smoothly integrate with popular short read aligners, genome browsers, and other tools to form a complete analysis pipeline. Components of PEGASUS have already been used for the analysis of allele-specific activity included in the integrative paper published by the ENCODE Project Consortium (Consortium et al., 2012b), reporting the results of the largest effort to date to identify functional elements in the human genome. PEGASUS is also presently being used for analysis of allele-specific activity in data produced by the Roadmap Epigenomics Mapping Consortium (REMC) (Kundaje et al., 2015).

An overview of a workflow for aligning reads to a personal genome and identifying allelic reads is shown in Figure 2-5. The first step in this process is to create the diploid genome sequences. PEGASUS splits the task of personal genome creation into two steps, assignment of variants to haplotypes and diploid sequence generation, that can be performed separately. Decoupling these two steps simplifies the haplotype assignment process and gives the user the ability to review the haplotype assignments of unphased variants and optionally use other tools to fine tune them, a feature which is lacking in existing software for creation of diploid genomes. It also allows the haplotype assignment process to be performed separately from the I/O intensive process of generation of the diploid sequences.

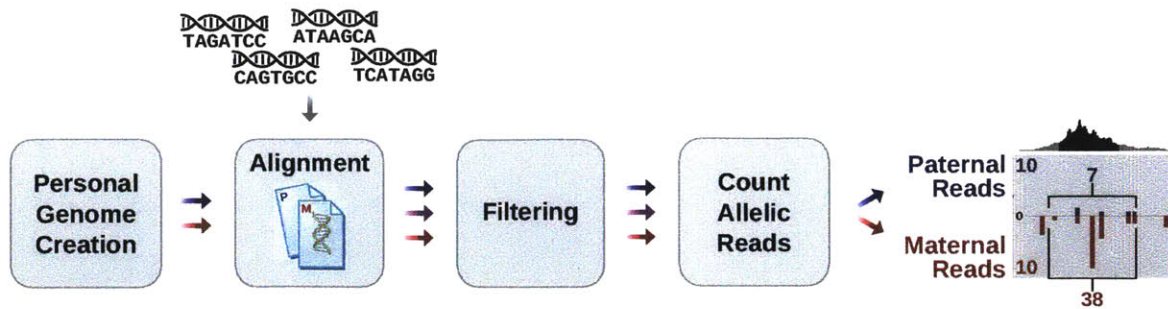


Figure 2-5: Overview of workflow for aligning reads to a personal genome and detecting allelic activity with PEGASUS. A personal genome is created and used for alignment of sequenced reads. After appropriate filtering steps are applied PEGASUS can identify reads overlapping variants and count the number of reads with each allele. If the variants are phased PEGASUS can take advantage of that information to detect allele-specific activity at functional elements such as ChIP-Seq peaks.

2.3 Haplotype assignment and creation of personal genomes

2.3.1 The simplest case: non-overlapping variants

The process of constructing a personal genome involves combining one or more sets of variant calls for the individual, or cell line, with a reference sequence to produce two separate, modified versions of the reference sequence containing the alleles of each variant and representing the actual diploid genome of the individual. As input, PEGASUS takes reference sequences in the (standard) FASTA format (Pearson and Lipman, 1988) and variant calls in the VCF format (Variant Call Format; Danecek et al., 2011). The first step in this process is haplotype assignment, the process of assigning the two alleles of each variant to the two homologous chromosomes of the diploid genome¹.

In an unrealistically simplified case, when the variants consist only of non-overlapping, and therefore non-conflicting sets of SNPs (it would be atypical for there to be conflicting SNP calls), and when there is no phasing information, then

¹We will often substitute the term “variant” for “variant call” for simplicity, even though technically a variant call is really an estimate of a variant that is believed to be in the sequence

| Reference Sequence | Variant call (chr:pos, ref, genotype) | Diploid sequence with variants assigned |
|--------------------|---------------------------------------|--|
| ...GGGGCTTC... | chr14:73393878 G G/A | Haplotype 1: ...GGGGCTTC... Haplotype 2: ...GGGACTTC... |

Figure 2-6: An unphased, heterozygous SNP with a reference allele of “G” and an alternate allele of “A” is incorporated into a personal genome by creating two copies of the reference sequence and substituting the alternate allele into the sequence for the haplotype to which it has been assigned, “Haplotype 2” in this example.

constructing a personal genome is rather straightforward. First, for each SNP, the alleles of the SNP indicated by the genotype of the individual can simply be assigned to two haplotypes, designated “haplotype one” and “haplotype two”, with the assignment chosen in a uniform random manner. Second, two copies of the reference genome can be made to represent the two haplotypes, with the SNP alleles substituted according to the haplotype assignments (Figure 2-6).

Although the previous example ignored phasing information, in practice it is highly preferable to have phasing information available when constructing personal genomes. As described previously, phasing information for variants may be generated in a local, relative manner through the process of read-backed phasing, or in a genome-wide, absolute manner through the process of trio-phasing. As before, when the variants are only SNPs, then satisfying the constraints resulting from the phasing information is also rather straightforward. Supposing there is local, read-backed phasing information, then instead of making a random assignment for each SNP, a random assignment is made for each phase set and applied to all of the SNPs in the set (Figure 2-7). Alternatively, if there is trio phasing information then the alleles of the SNP are assigned to the maternal and paternal haplotypes as indicated by the phasing information.

Extending the example to include indels (while maintaining the non-overlapping property of the variants) doesn’t change the haplotype assignment process, but alters the relative alignments of the sequences. Aligned portions of the diploid sequences will be offset from each other and the reference sequence

| Reference Sequence | Variant call (chr:pos, ref, genotype) | Phase Set | Diploid sequence with variants assigned |
|--------------------|---------------------------------------|-----------|---|
| ...GGGGCTTC... | chr14:73393878 G G/A | G A, C T | Haplotype 1: ...GGGGACTTC... |
| | chr14:73393881 T T/C | | Haplotype 2: ...GGGGCTCC... |

Figure 2-7: The heterozygous SNPs which have been phased with read-backed phasing are assigned to haplotypes by making selecting a haplotype assignment for the phase set and applying it to both variants. In this case the alleles listed first in the genotypes described by the phase set are assigned to Haplotype 2.

| Reference Sequence | Variant call (chr:pos, ref, genotype) | Diploid sequence with variants assigned |
|----------------------|---|--|
| ...ATATTCACTGGGTG... | chr1:53697036 TCACTG TCACTG/ATAAATAGGA | Haplotype 1: ...ATATATAAATAGGAGGTG... Haplotype 2: ...ATATTCACTGGGTG... |

Figure 2-8: When the personal genome includes alternate alleles of indels that are shorter or longer than the reference allele it is necessary to keep track of the length change so the sequences can be compared with each other.

because the lengths of the indel alleles often differ. It is critically important to keep track of the relative alignments of the genomic coordinates of the original reference sequence and the two haplotypes of the personal genome (Figure 2-8). This must be done so that the sequences, their annotated regions, and the effects being detected by short-read sequencing-based assays can be compared and analyzed. The UCSC LiftOver tool (Kuhn et al., 2012) serves as a convenient method for converting between genomic coordinate systems.

2.3.2 Challenges of haplotype assignment

Of course, real genomes contain SNPs, indels, long deletions, and other structural variants, and for several reasons, personal genome creation can be substantially more complicated when they are all included in the sets of variant calls. First, unlike the previous example, there will be overlapping variants. It is, naturally, a common occurrence for heterozygous SNPs and indels on one chromosome to be overlapped by longer indels and long deletions occurring on the homologous

chromosome. As described previously, however, different types of variants are detected by different types of assays and called by different, specialized algorithms, but all are reported relative to the reference sequence. As a result, the correct haplotype assignment is not indicated in the variant calls. This is most easily demonstrated by considering the case of long deletions (Figure 2-9). For example, the NA12878 genome includes 27 heterozygous long deletions, greater than 3kb long, detected by analysis of fosmid sequencing (Kidd et al., 2008, []). These long deletions overlap dozens (or hundreds in the case of the longest deletions) of SNPs and shorter indels detected by other methods. In this scenario a variant call for a heterozygous long deletion will indicate one allele with the sequence representing the deletion and a second allele with the exact sequence of the reference allele. The indels and SNPs that are overlapped by the long deletion can obviously still occur on the haplotype without the long deletion. Accordingly, the true sequence of that haplotype won't match the reference sequence indicated by the variant call for the long deletion. An additional complication is that the SNPs and shorter indels that are overlapped may be reported as either heterozygous or homozygous for the alternate allele, depending on the variant calling algorithm, even though the alternate allele could only possibly occur on one haplotype if the other haplotype is truly altered by the long deletion.

Second, as discussed previously, sequence variation can often be explained in multiple ways, for example as different combinations of SNPs and indels, that are inconsistent with each other. Since it is common practice for multiple variant calling algorithms to be used, it is, accordingly, common for these multiple, inconsistent, explanations of sequence variation to be reported. Furthermore, it is not even necessary for multiple variant callers to be used for this to happen, some variant calling algorithms will report multiple ways that a sequence variation could be explained and use the quality score to indicate the likelihood of each.

Finally, phasing information further constrains and complicates the task of resolving conflicting variant calls and making haplotype assignments. When trio-

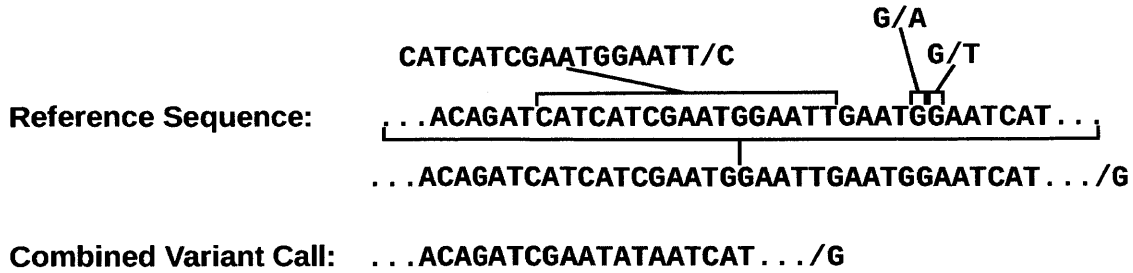


Figure 2-9: In this example the reference sequence (middle) is spanned by a heterozygous long deletion (just below reference) as reported by one variant caller. Meanwhile, other variant callers have reported a heterozygous indel (top left) and two heterozygous SNPs (top right). It is possible for the long deletion to have affected one haplotype while the other variants have their alternate alleles on the other haplotype. The alleles which are reported by the variant callers won't match this arrangement, however, because they are all reported relative to the reference sequence. The genotype for a "combined" variant call showing the actual alleles that occur when the alternate alleles are on opposite haplotypes is also shown (bottom).

phasing information is available, the overlaps must be resolved and the individual variants must be assigned consistent with the specified paternal and maternal haplotypes. When only read-backed phasing information is provided, the variants in a phase set are collectively constrained to be assigned to the same haplotype. Phase sets may span tens or hundreds of kilobases, and therefore a phase set containing variants from one source of variant calls may overlap multiple variants from another source of variant calls (that may or may not have phasing information). The haplotype assignments of these multiple sets of overlapping variants must be resolved together in order to find an assignment that satisfies all the phasing constraints. As in the case when there is no phasing information, there will be at least two equivalent assignments, with the haplotypes swapped, and one of them may be selected in a uniform random manner.

When resolving overlapping or conflicting variant calls, in the best case, when at most two called variants overlap in any location, both are heterozygous, and they aren't constrained to be on the same haplotype by phasing information, then the alternate alleles can simply be assigned to opposite haplotypes. If, however,

more than two called variants overlap, one or more are homozygous for the alternate allele, or phasing indicates alternate alleles are on the same haplotype, then the set of variants is overconstrained and one or more of the overlapping variants must be omitted.

One final task, which is critically important to facilitating downstream analysis, is to generate “combined” variant calls for overlapping variants (Figure 2-9), so that the variants may easily be analyzed together. This is important to prevent false positive signals of allele-specific activity, and a variety of double-counting types of errors, for example, and also facilitates comparing the locations of the variants with other genomic annotations.

2.3.3 Maximum-likelihood haplotype assignment using a context-sensitive input/output hidden Markov model

The overall challenge of haplotype assignment is to resolve the conflicts that can occur when called variants overlap while satisfying any constraints from phasing information. As described in the previous section, there are a variety of scenarios in which it is impossible to resolve conflicts in a manner that includes all the variants. Identifying an optimal haplotype assignment for conflicting variant calls requires considering all possible combinations of assignments of those variants. One of the contributions of this thesis is a method for generating a maximum likelihood haplotype assignment of variants based on the quality scores of the variant calls. This is accomplished by combining two extensions of the standard HMM that have previously been described in the literature, a Context-Sensitive HMM and an Input-Output HMM (Yoon and Vaidyanathan, 2004; Bengio and Frasconi, 1995). The novel formulation of an HMM that is produced by combining these is a Context-Sensitive Input/Output Hidden Markov Model (csIOHMM), and it is a natural fit for modeling the haplotype assignment process. The task of considering all the possible combinations of variant assignments is made computationally tractable and performed efficiently by taking advantage of dynamic programming

and using a modified version of a Viterbi decoding, the algorithm used for determining a maximum likelihood state sequence for standard HMMs.

A graphical model diagram representing the csIOHMM is shown in Figure 2-10. In this application the state sequence corresponds to the bases of the reference genome. In general, all of the information that would be encoded in the hidden states of the HMM is instead provided by the input and by the memory². The implementation of the csIOHMM in PEGASUS operates with an initial start state, a single context-sensitive run state, which it stays in for the complete length of the genomic sequence, and a final, terminal state. The input for the csIOHMM, corresponding to each node in the state sequence, are the variant calls, if any, that begin at that genomic coordinate of the reference sequence. Indels and local phasing information may affect multiple sequence positions, which necessitates the HMM being context-sensitive, and these effects are accounted for using the memory associated with each sequence position. When haplotype assignments are made for deletions the memory for the positions with deleted bases are updated to indicate that. Similarly, when a haplotype assignment is made for a phase set that can be indicated in the memory for all the other positions at which there are variant calls that are part of the same phase set. The output produced for each position in the state sequence is the result of the haplotype assignment of the variant calls in the input, if any. When variants conflict with each other such that one or more must be omitted, a maximum-likelihood assignment is made by using the quality scores of the variant calls. If quality scores are not directly comparable between sets of variant calls then the modified Viterbi decoding can instead maximize the quality scores separately for each source of included variants. If quality scores are unavailable for some (or all) variants then the total number of variants that are included from the source (or from all sources) can be maximized instead.

In a standard HMM the maximum-likelihood state sequence can be efficiently

²The csIOHMM model does, of course, allow for a variety of ways that the set of available states could be defined. For example the csIOHMM could be formulated with a state corresponding to each source of variant calls, in which case the current state indicates the set of variant calls that best explain the variation seen in the current region of the genome

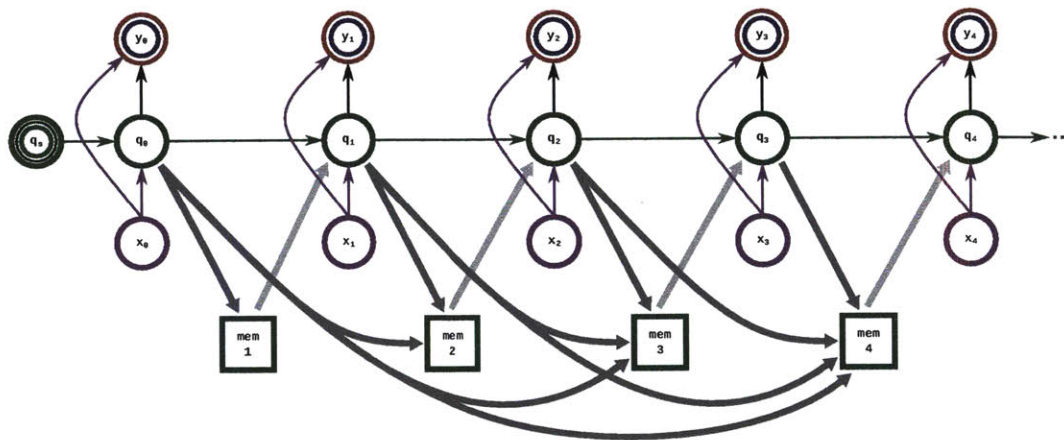


Figure 2-10: A graphical model diagram for a context-sensitive Input/Output Hidden Markov Model. The state sequence “q” (green circles) corresponds to the positions of the reference sequence. At each position the input “x” (purple circles) is a set of all variant calls for that position. The context stored in memory at each position (green boxes) indicates whether the reference base at the position has been altered by a previous variant along with any haplotype assignments that have been made for phase sets containing variants at the position. The output “y” at each position (blue/red double circles) is a set of variant calls corresponding to the input with haplotype assignments for the alleles indicated, or an indication that the variant call is omitted.

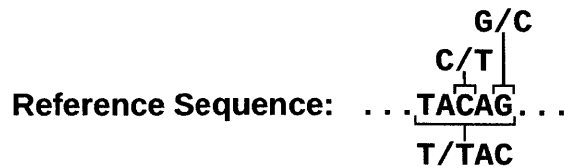


Figure 2-11: In this example an indel with a heterozygous genotype specifying two alternate alleles (bottom) overlaps two SNP calls (top). The G/C SNP overlaps a position that is deleted in both of the indel alleles, therefore one of these variant calls must be omitted from the personal genome. At the same time, the C/T SNP occurs at a position that is unaltered by the TAC allele of the indel. Although it will always be possible to make a valid haplotype assignment for that SNP, the possible choices for haplotype assignments will depend on the haplotype assignment of the indel, or whether it is omitted.

determined using the Viterbi decoding, which is a dynamic programming algorithm. The standard Viterbi decoding computes optimal sequences of states and extends them incrementally. This is possible because of the Markov property; the future state of the HMM depends only on the current state. This doesn't work for a context-sensitive HMM, however, because the context can create dependencies that may span a long range of the observed data. In the case of haplotype assignment, these dependencies result from deletions and local phasing of variants. More specifically, in scenarios in which sets of overlapping variants are overconstrained, such that one or more must be omitted, the haplotype assignments of all of those variants will depend on each other. Furthermore, there may be other variants that overlap just one of the variants in the set. Although it may always be possible to find a haplotype assignment for those variants their haplotype assignment will still depend on the haplotype assignment for the overconstrained variant that they overlap (Figure 2-11).

As well, in the case of phase sets, the haplotype assignments made for all the variants that are members of a phase set will depend on each other even though other variants that aren't part of the phase set may be interspersed among them. Not surprisingly, the most complicated scenarios arise when the variants in a phase set overlap other multiple sets of other overlapping variants. Despite

the existence of these dependencies, due to their structure it is still possible for optimal subsequences to exist, and dynamic programming can still be used to efficiently determine optimal haplotype assignments.

The modified Viterbi decoding used for haplotype assignment differs from the standard dynamic programming algorithm by splitting up the variant calls in order to find sets for which optimal subsequences can be computed. Haplotype assignments are made independently for each chromosome. For each chromosome the algorithm iterates over three steps: First, a "current" set of variants that may need to be assigned together is generated. Second, overlapping variants are detected and the subsets for which there will be optimal subsequences are identified. Third, a maximum-likelihood haplotype assignment is computed.

In the first step, the algorithm makes a complete pass over the input variants to find variants that belong to phase sets and identify the first and last variant of every phase set. Next, as variants are read in unison from all the sources in order of genomic coordinates the algorithm checks whether they belong to a phase set and conducts a simple check for overlap with other variants based on the start coordinates and lengths of the reference alleles. When a variant doesn't overlap any others and is neither in a phase set nor within the region spanned by a phase set then a haplotype assignment can be made for the individual variant and the process can start over for the next variant. When overlapping variants or variants belonging to a phase set are found the algorithm continues reading variants, building up a collection, until it has both reached the end of all phase sets containing variants in the collection and the next variant doesn't overlap any variants in the collection.

After a collection of variants has been assembled the second phase of the algorithm performs a more detailed check for overlaps that excludes partial overlaps with indels that occur when the only portion of the indel that is overlapped is part of a common prefix or suffix that is shared by both of the genotyped alleles and the reference allele. As overlapping variants are found the algorithm can also evaluate whether they are overconstrained for any of the reasons described previ-

ously in section 2.3.2. Although making the assignments for the overconstrained variants will require additional computation, the algorithm will be able to compute optimal subsequences of haplotype assignments for the variants that are not overconstrained.

Having accurately identified sets of overlapping variants the next step is to determine whether there are any phase sets that have variants in multiple of the sets of overlapping variants. If so, the haplotype assignments for those sets will be dependent on each other. In order to handle that scenario the algorithm will compute the maximum-likelihood haplotype assignments for all the overlapping sets of variants for all the combinations of haplotype assignments of the phase sets and then select the combination of haplotype assignments for the phase sets that results in the overall maximum-likelihood haplotype assignment. If there are no overlapping variants among the current collection then haplotype assignments may be made independently for each entire phase set, and each individual variant that is not a member of a phase set.

In order to efficiently determine the maximum-likelihood haplotype assignment for individual sets of overlapping variants the algorithm makes use of the knowledge of which variants are overconstrained (i.e. one or more will need to be omitted). Outside the ranges of overconstrained variants the optimal state sequence can be computed using a traditional dynamic programming approach. The algorithm can keep track of two optimal haplotype assignments for the previous variant: alleles assigned in the order indicated by the genotype, and the reverse order³. For each of those same two orderings the optimal haplotype assignment up through and including the current variant will be the best combination of that haplotype assignment with an optimal haplotype assignment up through the previous variant.

For the variants shown in Figure 2-9, for example, the algorithm would first make a haplotype assignment for the long deletion, keeping track of the bases

³if a variant were homozygous the choice is simply whether or not to omit it, but it could not be part of a phase set, and would be overconstrained if it overlapped any other variants

that have been deleted on the haplotype with the alternate allele. For the first variant it's possible to optimize by assigning the alleles to the haplotypes in the order specified in the genotype, because the variants are not overconstrained and none are trio-phased. Next, the algorithm checks both possible assignments for the indel. Only the assignment that places its alternate allele on the haplotype without the deletion will be valid. The same steps will then be taken consecutively for each SNP, and likewise the only valid haplotype assignment for each will be for the alternate allele to be on the haplotype without the long deletion. Finally, if there is no trio-phasing information for the variants, then the haplotype assignments can always be swapped to produce an assignment with an equivalent score, so a choice can be made at random to either keep the haplotype assignment as is, or swap all the assignments.

Within the ranges spanned by overconstrained variants it's necessary for the algorithm to keep track of all the valid combinations of haplotype assignments for overconstrained variants, but memoization can still be used to avoid repeating calculations. For each overconstrained variant the possible haplotype assignments will include the choice to omit the variant in addition to both possible assignments of the alleles to haplotypes (or just one possible assignment if the variant is homozygous). There may also be variants that overlap an overconstrained variant but are not overconstrained. While computing an optimal haplotype assignment for the non-overconstrained variants the algorithm must continue to keep track of the possible combinations of haplotype assignments for the overconstrained variants, but can keep track of only the optimal haplotype assignments for the non-overconstrained variants.

When quality scores are available for all the variant calls and are comparable across all sources of variant calls then the likelihood is calculated as the sum of the phred-scaled (logarithmic) scores of the variants that are omitted. The scores represent the negative log probability that the variant call is incorrect, so better assignments will have lower likelihood scores. If quality scores are not comparable across sources of variant calls they can still be used to compare assignments

of variant calls from the same set, but priorities for the different sets of variant calls must be provided as parameters to the algorithm. When quality scores are not available or priorities haven't been provided the algorithm instead seeks to minimize the number of variants that are omitted. If the sets of variants are prioritized this can be done separately for each set, otherwise it is done collectively for all the variants.

PEGASUS also improves upon existing personal genome creation tools in two additional ways. First, priority levels can be assigned to input files containing the variants for the purpose of resolving conflicts between overlapping variants. When priority levels are specified the maximum-likelihood (or number of included variants if quality scores are unavailable) is tracked separately for each set of variants, and variants with a higher priority level are always favored, such that the haplotype assignment of the highest priority variants will have a maximum-likelihood assignment based solely on the algorithm's ability to include those variants in the personal genome. The haplotype assignments of lower priority variants will then be maximized subject to the algorithm's ability to include those that don't conflict with any higher priority variants. This is particularly useful when variant calls have quality scores that are not directly comparable, or were produced using multiple assays or analysis methods with different false-positive rates. Second, PEGASUS produces a master output VCF file in which each set of overlapping variant calls have been combined into a single variant call with alleles representing the actual sequences found in the personal genome, as shown in Figure 2-9.

2.3.4 Personal genome creation

Once the conflicts have been resolved and the haplotype assignments have been made, the process of generating the diploid sequences from the variant calls is rather straightforward. The personal genome creator component of PEGASUS generates the two sequences of a personal, diploid genome from a reference se-

quence by incorporating variant calls according to their haplotype assignments. PEGASUS reads each chromosome of the reference sequence and the haplotype assignments for the variants on that chromosome in an iterative manner. The portion of a chromosome preceding the first variant is simply output unchanged to each haplotype sequence. Then either the sequence of the reference allele or an alternate allele is output to each haplotype sequence according to the assignment of the first variant. These same two steps are then repeated for the entirety of each chromosome.

As each variant is processed, the personal genome creator also generates an updated, final master VCF file that in addition to containing combined alleles of overlapping variants and the phasing that was applied for every variant, indicates the coordinates of each variant in both of the diploid sequences. The master VCF file greatly facilitates downstream analysis and to our knowledge there are no other tools that produce one. As previously mentioned, for indels and structural variants PEGASUS keeps track of any differences in length among the haplotype sequences and the reference sequence. Any such differences are reported in UCSC LiftOver chain files (Hinrichs et al., 2006) that map coordinates in both directions between the two haplotype sequences and between each of the haplotype sequences and the reference.

2.4 Comparison of personal genomes created with PEGASUS and AlleleSeq

We compared PEGASUS with AlleleSeq (Rozowsky et al., 2011) by using each tool to create personal genomes for two cell lines, H9, and STL001, that were selected because of their inclusion in the datasets produced by the Roadmap Epigenomics Mapping Consortia. Whole genome sequencing of the H9 and STL001 cell lines was completed by the REMC. We collaborated with researchers at Baylor College of Medicine to produce consensus variant calls that combined the output of sev-

| H9 cell line | | |
|-------------------------------|-----------|-------------|
| | SNPs | indels |
| Total number of variant calls | 3,645,544 | 589,189 |
| Number of overlapping sets | 5,167 | |
| Number omitted by AlleleSeq | 0 | 2,143 (41%) |
| Number omitted by PEGASUS | 0 | 949 (18%) |

(a)

| STL001 cell line | | |
|-------------------------------|-----------|-------------|
| | SNPs | indels |
| Total number of variant calls | 2,968,433 | 595,232 |
| Number of overlapping sets | 2,905 | |
| Number omitted by AlleleSeq | 0 | 1,312 (45%) |
| Number omitted by PEGASUS | 0 | 653 (22%) |

(b)

Table 2-1: Each table shows the total number of variant calls for one of the cell lines that were used to test PEGASUS and AlleleSeq, followed by the number of sets of overlapping variants, and the number of variants there were omitted from the personal genomes by each utility. AlleleSeq consistently omits more than twice as many variants as PEGASUS because AlleleSeq fails to find valid haplotype assignments for them.

eral variant calling programs, and to perform read-backed phasing for the SNPs. The H9 variant calls included more than 3.6 million SNPs and nearly 600,000 indels. The STL001 variant calls included nearly 3 million SNPs and nearly 600,000 indels. These variant calls and the number that were omitted from the personal genomes by PEGASUS and AlleleSeq are summarized in Table 2-1.

Chapter 3

Methods for analyzing sequenced reads with personal genomes and for detecting allele-specific activity

This chapter presents additional methods included in the PEGASUS software for processing reads that have been aligned to a personal genome and for detecting allele-specific activity. We discuss how separating the reads that overlap variants from those that don't overlap variants facilitates efficient processing of the reads, as well as a method for filtering PCR duplicates in a variant-aware manner. We also describe methods for accurately measuring allele-specific activity at individual variants and present the results of a comparison between PEGASUS and AlleleSeq for detecting allele-specific activity. Finally, we demonstrate how we can take advantage of variant phasing information to detect allele-specific activity across regions with greater sensitivity than at individual variants.

3.1 Introduction

In any single cell at a specific moment in time it is possible that the transcription of a particular gene might be occurring on one chromosome, but not on the homologous chromosome. If there is a heterozygous variant in the coding sequence

of that gene then this can be detected by performing single-cell RNA-Seq and examining the aligned reads that overlap that variant; we would see that all the reads would contain the same allele of the variant. If the allele-specific expression detected in the single cell occurred by chance then we would expect that when we examine the reads from an RNA-Seq assay performed on millions of cells that the reads overlapping the variant would contain both alleles in approximately equal numbers. If, however, there is a biological reason for the cell to be transcribing the gene exclusively, or primarily, from one chromosome, such as a sequence variant in the gene's regulatory region, then we would expect to see that many more reads should have one allele than the other. Allele-specific binding of transcription factors and allele-specific histone modifications can be detected in an analogous manner by examining reads from a ChIP-Seq assay.

Allele-specific activity can provide valuable information about the way cells function, and the functional effects of sequence variants. Alignment of sequenced reads is only the first step in the analysis of sequencing based assays, however. Before the aligned reads can be used for detecting allele-specific activity, or comparing levels of gene expression, or calling ChIP-Seq binding peaks, there are typically other processing steps that must be completed. Although the use of a personal genome allows for a more accurate and complete alignment of sequenced reads than a reference genome, tools for aligning and processing sequenced reads are not typically designed to work with a personal genome, or with reads that have been aligned to a personal, diploid genome sequence.. In addition to tools for haplotype assignment and personal genome creation, PEGASUS facilitates using reads that have been aligned to a personal genome with the software tools that are used in typical analysis workflows for sequencing based assays, and includes tools for detecting allele-specific activity.

3.2 Incorporating a personal genome into standard workflows

In the case of alignment, for example, aligning the reads to both haplotypes at once would result in all of the reads that don't overlap variants mapping equally well to at least two locations in the personal genome, because the two haplotypes will be identical where there are no variants. Programs for read alignment typically have parameters that enable reporting of non-uniquely aligned reads, but specifying those parameters may not produce the intended results. For example, the aligner would report reads that align non-uniquely due to homology in the genome or sequencing errors that would otherwise be suppressed, and these reads would be difficult to distinguish from the reads that map to equivalent locations on the two haplotypes.

Instead, the workflow that PEGASUS supports is to align the sequenced reads to each haplotype separately. This allows the user to choose the aligner that is best suited to aligning the type of reads that need to be processed. The two sets of aligned reads that are produced can then be processed together by PEGASUS to separate reads that overlap variants from those that don't. For reads that overlap variants and align to both haplotypes the mapping quality of the bases that overlap the variant, and the mapping quality of the read as a whole (Li et al., 2008a) are compared for each haplotype and PEGASUS will report the haplotype that the read maps to the best. Alternatively, a tie is reported when a read maps equally well to both haplotypes (this can happen when a read overlaps indels, or when the read is aligned to different locations in the two haplotypes, for example). Separating the reads that overlap variants from those that don't also allows for downstream analyses that only involve reads that overlap variants, such as detection of allele-specific activity, to be performed much more efficiently, because the vast majority of reads produced in sequencing based assays won't overlap variants and the I/O time overhead of reading them can be significant.

After the reads that overlap variants have been separated from those that don't

the next step is typically marking or removal of PCR duplicates. The reads that don't overlap variants may be processed with a standard tool while those reads that do overlap variants can be processed using a variant-aware PCR duplicate removal utility that is part of PEGASUS. Additional steps in the analysis workflow, peak calling for example, can be performed separately for each haplotype using the PCR-duplicate filtered reads. Alternatively, by using the UCSC Liftover files that were generated by the personal genome creator and the PEGASUS Liftover-SAM utility, the reads can be mapped back to the reference genome's coordinate system for the same steps to be performed.

3.3 Variant-aware detection of PCR duplicates

Removal of PCR artifacts is an important quality control measure for short read sequencing assays. Failure to remove them can easily lead to errors when measuring gene expression, calling ChIP-Seq peaks, or detecting allele-specific activity. The simple approach taken by utilities such as samtools (Li et al., 2009) is to classify all reads with the same 5' start coordinate as duplicates and retain only the single read with the highest quality score. In general this can be problematic with sufficiently deep sequencing because it becomes expected that there will be more than one read with the same start coordinate. It is even more of a problem when working with personal genomes and measuring allele-specific activity. Existing tools won't distinguish between the two haplotypes and will mark reads as PCR duplicates even if they contain different alleles and therefore originated from different haplotypes. This may bias the allelic read count towards the allele in the single chosen read, and the result can be significant when it affects the same variant multiple times, or when the read counts are aggregated for phased variants within a genomic region.

PEGASUS includes a novel variant-aware duplicate marking utility¹. By split-

¹The PCR duplicate detection algorithm in PEGASUS also works for reads that do not overlap variants.

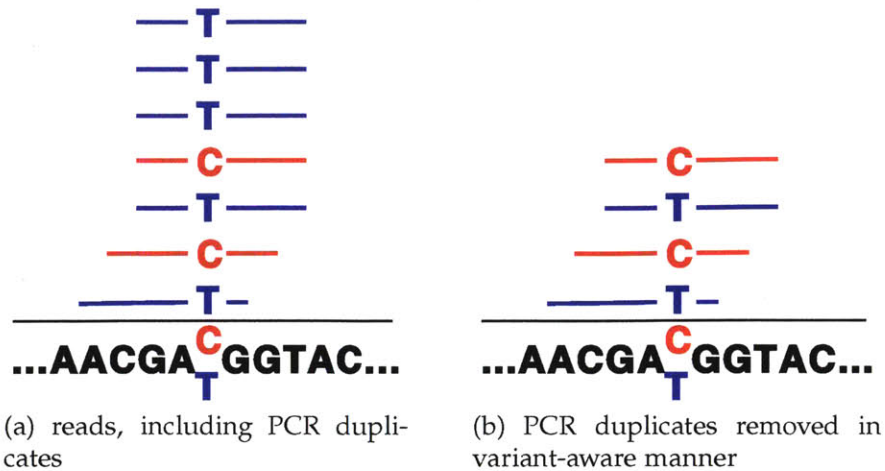


Figure 3-1: (3-1a) Reads that have the same start coordinate (and gaps, if any) are identified so they can be filtered if they are PCR duplicates. Four reads with a “T” allele and one read with a “C” allele have the same start coordinate. Standard PCR duplicate marking utilities would not distinguish between these reads and would pick one to keep and mark the rest as duplicates. (3-1b) PEGASUS recognizes that reads with different alleles are, in fact, not PCR duplicates, and keeps one read with each allele while removing (marking) duplicates.

ting reads that have the same start position into groups according to the bases overlapping variants PEGASUS is able to mark (or remove) PCR duplicates separately for each allele (Figure reffig:pegasus-other:pcr-dup-removal). At most one read for each allele is retained by default, but when sequencing is sufficiently deep that duplicate reads are expected a file with a maximum coverage at each variant may be provided as input to enable PEGASUS to keep up to a specified maximum number of reads instead.

3.4 Detecting allele-specific activity at heterozygous variants

In order to detect allele-specific activity at the location of a particular variant it’s necessary for the variant to be heterozygous and the number of aligned reads that include each of the alleles must be counted accurately. There are several sources of error that must be accounted for in order to get accurate allelic reads counts. The

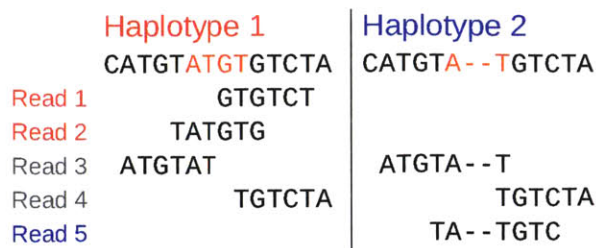


Figure 3-2: When two alleles of an indel share a common prefix or suffix a read that partially overlaps the indel can align to both haplotypes. For example, a portion of haplotype 1 consists of fourteen bases containing an indel with reference allele “ATGT” (highlighted in orange) while the homologous sequence of haplotype-2 contains the alternate allele “AT”. Read 1 and Read 2 align uniquely to Haplotype 1 and Read 5 aligns uniquely to Haplotype 2. Read 3 and Read 4, however, only partially overlap the indel and align equally well to both haplotypes. PEGASUS detects these reads and categorizes them as aligning to both genomes rather than including them in the counts of allelic reads.

most basic of these is that it is possible for reads from one haplotype to align to the other haplotype with mismatches at variant positions. This can occur even when the short reads are aligned to the diploid genome sequences separately and with parameters that strictly limit the number of allowed mismatches. Although some of these errors may be detected when reads that overlap variants are separated from those that don't, when counting allelic reads PEGASUS explicitly checks that the bases overlapping the variants match the allele that has been assigned to the haplotype to which the read mapped. In the case of SNPs PEGASUS requires that the base in the sequenced read match the allele exactly, and also accepts a parameter specifying a minimum quality score for the base in order for the read to count. For indels, PEGASUS does not require a perfect match, but uses mapping quality to determine which allele the read matches best, and also accepts parameters for specifying thresholds on the quality scores of both the matching bases and any mismatching bases. For example, reads that partially overlap indels and match exactly to the partial sequence of both alleles also will not get counted toward either allele (Figure 3-2).

A second possible source of error is PCR duplicates, which are best handled before the allelic reads are counted, as has already been discussed. A third po-

tential miscounting error can occur when reads align to overlapping variants. For example, consider a heterozygous SNP and a heterozygous indel that overlap and have their alternate alleles assigned to opposite haplotypes. If the variants are handled separately, then as the reads are counted for both variants there will be no reads that match each variant's reference allele. Instead, PEGASUS uses the alleles reported in the combined variant call in the master VCF file so the number of reads with the SNP's alternate allele can be directly compared to the number of reads containing the indel's alternate allele.

A final source of error are differences between the two haplotypes in the number of k-mer subsequences around a variant that are unique in the haplotype. The number of unique k-mer sequences that overlap a variant can differ between alleles due to k-mers containing one allele occurring at other locations in the genome. Therefore, it is an important to account for mappability differences to prevent false positive signals for allelic events caused by an inability of some reads to align uniquely to one haplotype. In the ENCODE uniform processing pipeline differences in mappability were corrected for by creating uniqueness maps for the hg19 reference sequence and excluding reads that aligned to non-unique locations (Hoffman et al., 2012). The same method can be used to create uniqueness maps for each haplotype of a personal genome. PEGASUS adjusts for differences in the number of unique positions around a variant by scaling the read counts by the ratio of the possible maximum number of unique k-mers to the actual number of unique k-mers overlapping each variant. This gives an estimate of the number of reads that would have been expected to have aligned if all the k-mers were unique. PEGASUS also accepts a parameter to specify a threshold on the mappability difference between alleles; allelic read counts will not be reported for a variant if the difference exceeds the specified threshold.

Once accurate allelic read counts have been measured we determine whether there is allele-specific activity at the location of a variant by testing a null hypothesis that the reads that overlap the variant are equally likely to have come from each haplotype. This follows from an assumption that the biological event that

was assayed is equally likely to occur on each haplotype, i.e. it is assumed that expressed genes are transcribed equally from both homologous chromosomes, and proteins are equally likely to bind to both homologous chromosomes. A binomial test can then be used to check whether a difference in the number of reads aligned to each haplotype is more extreme than would be expected under the null hypothesis. When simply checking for the occurrence of allele-specific activity a two-tailed binomial test is used. If, however, the difference in read counts is expected to be skewed towards a particular haplotype then a one-tailed test is used. Typically, we wish to check for allele-specific activity at many variants, so a Benjamini-Hochberg correction is applied to account for the multiple comparisons. A threshold on the minimum number of reads overlapping a variant may also be applied to reduce the number of comparisons that are performed and therefore reduce the magnitude of the multiple hypothesis correction.

3.5 Comparison of PEGASUS and AlleleSeq for detecting allele-specific activity

As part of an ongoing study of epigenomic allele-specific activity using data produced by the Roadmap Epigenomics Mapping Consortium, collaborators from the Milosavljevic Lab at Baylor College of Medicine used both PEGASUS and AlleleSeq to generate personal genomes for the H9 and STL001 cell lines. They then aligned CHIP-Seq data for six different histone modifications to the personal genomes and used both PEGASUS and AlleleSeq to measure allele-specific activity at variants in the H9 genome. We were provided with the aligned reads and allelic read counts produced by PEGASUS and AlleleSeq for comparison. Only measurements of allele-specific activity at SNPs were compared, because although PEGASUS detects allele-specific activity at indels, AlleleSeq does not.

Initially, it appeared that the total number of allelic reads at each SNP was consistently slightly higher for AlleleSeq. We determined that these differences

were the result of AlleleSeq not checking the quality scores of the bases that overlapped the SNPs, while PEGASUS requiring a minimum quality score for those bases. When the data were reprocessed using PEGASUS without a threshold on the quality score the differences were almost completely eliminated. The remaining differences in counts of aligned reads were attributed to differences in the personal genomes resulting from the stochastic manner by which both PEGASUS and AlleleSeq assign variants to haplotypes when they are not constrained by phasing information. More specifically, when the random assignments of alleles of nearby, unphased variants to haplotypes results in different configuration of alleles in the personal genomes produced by PEGASUS and Alleleseq, then reads that overlap both variants may align to one personal genome and not the other.

3.6 Detecting allele-specific activity at functional elements

Although individual variants may be examined to check for allele-specific activity, in some cases it may be of greater interest to determine the level of allele-specific activity across regions, or the read depth may be too low to reliably detect allele-specific activity at individual variants. If phasing information is available for the variants then PEGASUS can take advantage of that information to combine the signal from multiple variants and detect allele-specific activity across regions. In addition to boosting the ability to detect allele-specific activity when read depth is low, checking for allele-specific activity across regions results in fewer tests and therefore a less severe multiple hypothesis correction is necessary. Although PEGASUS does not impose any restrictions on the size of the regions used to check for allele-specific activity, because of the noisy nature of assays such as ChIP-Seq the best results are likely to be obtained when the regions are relatively small functional elements, such as ChIP-Seq peaks for the same protein for which allele-specific activity is being examined.

In order to check for allele-specific activity across a region we first compute the allelic read counts for all of the phased variants in the region. In the example shown in Figure 3-3 the allelic read counts range from 0 to 4 for all of the 26 SNPs in the region (23 of which are trio-phased). Next, the sums of the allelic read counts are computed for the maternal haplotype and (separately) the paternal haplotype, in the case of trio-phased variants, or for "haplotype one" and "haplotype two" for a set of read-backed phased variants. While computing the sum of allelic read counts, care must be taken to avoid double-counting reads that overlap multiple variants. PEGASUS is able to accomplish this because unlike other software the unique IDs of the reads are included in the metadata associated with the allelic read counts for each variant. Just as in the case of detecting allele-specific activity at individual variants, a binomial test can be used to determine whether the difference in allelic read counts is statistically significant. The data shown in Figure 3-3 is from the region of a ChIP-Seq peak for the histone modification H3K79me2 overlapping the TSS of the gene NACC2. Although the allelic read counts for all the individual variants were so low that they couldn't be statistically significant on their own, after taking advantage of the phasing information and aggregating across the region the binomial test produces a p-value of 1.16×10^{-4} and it can be determined that the histone modification is occurring primarily on the paternal chromosome.

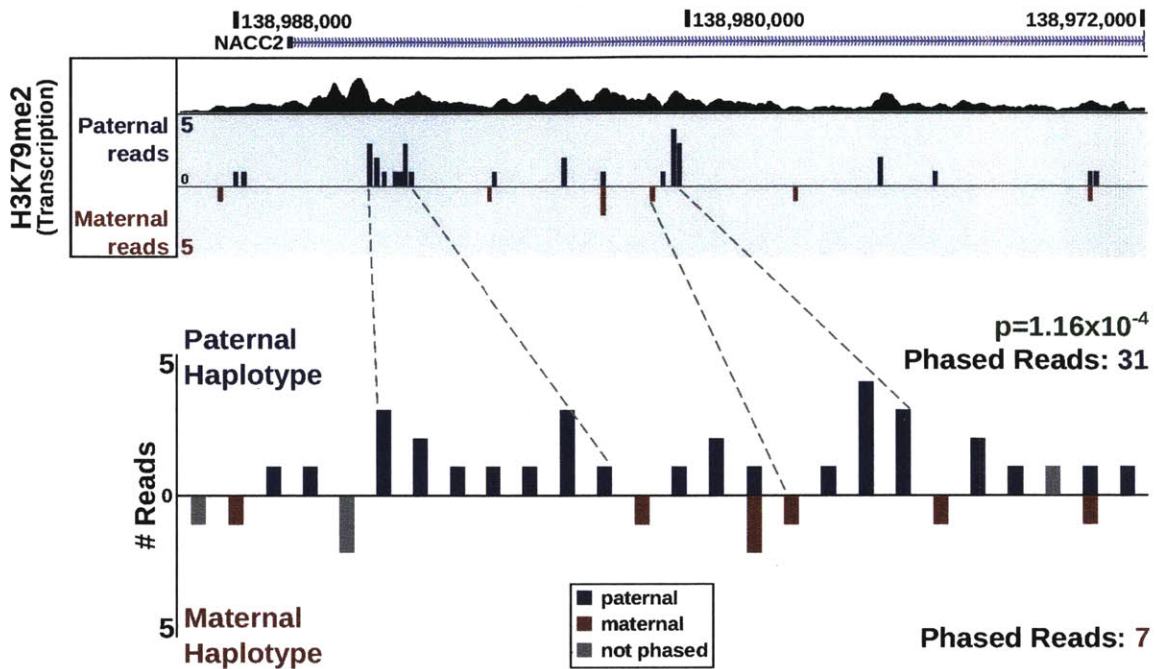


Figure 3-3: Allele-specific activity is detected at an H3K79me2 ChIP-Seq peak overlapping the TSS of the gene NACC2. Allelic counts are aggregated for 23 trio-phased heterozygous variants within the ChIP-Seq peak, while 3 unphased variants are omitted. The read counts at individual variants are all in the range from 0–4, and none would be statistically significant on its own. When aggregated there are 31 reads aligned to the paternal haplotype and 7 aligned to the maternal haplotype. The resulting p-value is 1.16×10^{-4} , indicating that the H3K79me2 histone modification is occurring primarily on the paternal chromosome.

Chapter 4

A genome-wide survey of allele-specific activity in a human genome

In this chapter we describe the analysis that was performed on data for the GM12878 cell line produced by the ENCODE Project Consortium. We examine allele-specific activity with a genome-wide perspective. We detect correlations in whole-genome measurements of allele-specific activity and demonstrate that we can gain insights into gene regulation by considering multiple signals of allele-specific activity. Results from this analysis were included in the ENCODE integrative analysis (ENCODE Consortium et al., 2012b)

4.1 Introduction

The ENCODE Project's goal is to identify all functional elements in the human genome. While many regions can be annotated as being likely to have function based directly on transcription, or the presence of ChIP-Seq peaks, or less directly from DNase hypersensitivity, detecting the occurrence of allele-specific activity can provide insight into the role of functional elements and the effects of sequence variation. The data produced by ENCODE for the GM12878 cell line were

an excellent resource for investigating this topic for several reasons: a wide variety of sequencing based assays were performed because GM12878 was a “tier 1” cell line; the cell line was derived from the daughter of a parent-child trio whose genomes were deeply sequenced in the pilot of the 1000 Genomes project (Consortium et al., 2012a) and high quality trio-phased variant calls were produced; GM12878 has been studied extensively previously because it is also one of the original HapMap cell lines (Gibbs et al., 2003).

4.2 Methods

At the time we began studying allele-specific activity in the ENCODE data, PEGASUS did not yet exist, and consortium members from the Gerstein Lab had already produced a diploid genome sequence for GM12878 using AlleleSeq. This genome incorporated 3,657,082 SNPs, 328,528 indels, and 1,544 long deletions, of which 2,247,802, 202,141, and 1,110 were heterozygous, respectively. In our initial analysis, however, we were limited to including just 1,409,992 phased, heterozygous SNPs, and 167,096 phased, heterozygous indels for which the haplotype assignments were readily verifiable, because AlleleSeq did not produce an output file that clearly indicated which variants were included, or haplotype assignments, or overlapping variants. After PEGASUS was developed a full accounting of the variants included in the personal genome was produced and the analysis was repeated using the complete set of heterozygous variants and with an updated processing pipeline. Although minor differences were found at specific genomic locations, because of the genome-wide nature of the analysis there were no significant differences in the results from the updated analysis. The available data included ChIP-Seq for 72 transcription factors, 11 histone modifications, RNA Polymerase II (POLR2A) and III, DNase hypersensitivity, and RNA-Seq assays performed using a variety of protocols (Consortium, 2011). All of these assays had multiple replicates (in some cases performed at different facilities).

For all the assays the sequenced reads were first aligned to the EBV genome

and hg19 mitochondrial chromosome sequences. Only those reads that did not align to the EBV and chrM sequences were subsequently aligned, separately, to each haplotype of the GM12878 personal genome. For ChIP-Seq data the Bowtie aligner was used for all alignment steps (Langmead et al., 2009), and for RNA-Seq data the TopHat aligner was used (Trapnell et al., 2009). PCR duplicates were removed using the variant-aware duplicate marking algorithm of PEGASUS. Allelic read counts were then computed for each assay, and counts were summed across replicates generated by the same facility. Variants were excluded if they overlapped regions that had been blacklisted by ENCODE due to poor mappability. For all variants, read counts were adjusted to correct for any differences in the mappability of the maternal and paternal alleles. Phasing information was used to combine allelic read counts and measure allele-specific activity for three types of regions: ChIP-Seq peaks, chromatin state segments identified by ChromHMM (Hoffman et al., 2012; Ernst and Kellis, 2012), and gene bodies. Rather than generating new ChIP-Seq peak calls based on the reads aligned to the personal genome we chose to use the ChIP-Seq peak calls generated by the consortium, because an extensive sequence of quality control procedures had already been applied to them.

4.3 Validation of method for detecting allele-specific activity

For the purpose of validating our methods we took advantage of a unique feature of the GM12878 cell line, biased X-inactivation, and compared allele-specific activity on the X chromosome to published results showing the frequency with which genes escape X-inactivation. It has been shown that the paternal X chromosome is inactivated in approximately 90

The overall allele-specific activity in the vicinity of the ChIP-Seq peaks called by ENCODE closely mirrored the pattern of escape from X-inactivation reported

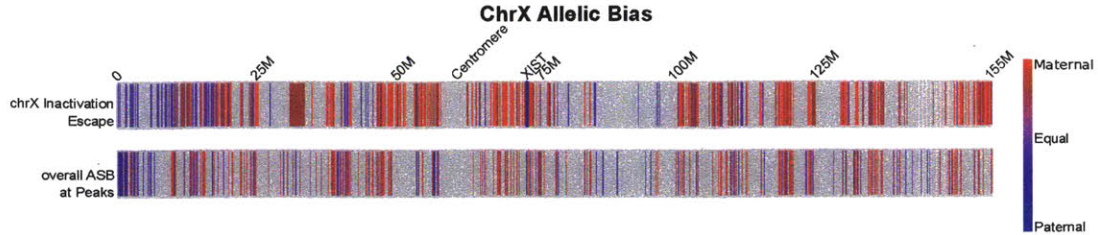


Figure 4-1: Comparison between frequency with which genes escape from X-inactivation and aggregate allele-specific activity of POLR2A, activation associated histone modifications, and TFs known to be activators as measured at ChIP-Seq peaks on the X chromosome

in (Carrel and Willard, 2005) (Figure 4-1). In order to ensure that we didn't confound the allele-specific activity due to X-inactivation with biases due to allelic binding of a transcription factor we excluded peaks with disrupted motifs. We found an R^2 of 0.73 (Pearson's) for the correlation between frequency of escape from X-inactivation and average allelic bias for POLR2A and a set of transcription factors known to be activators.

Although it would have been appealing to use imprinted genes to validate measurements of allele-specific activity, they were, in fact, impractical for a variety of reasons: Relatively few imprinted genes are known, imprinting only occurs in specific cell types, and limited evidence is available to support the imprinted behavior of genes included in catalogs of imprinted genes.

4.4 Genome-wide allelic correlations

Given the genome-wide scope of ENCODE we began by surveying whole genome correlations in measurements of allele-specific activity in 193 ENCODE assays. For each pair of signals correlations were computed for allele-specific activity within regions of both gene bodies (below the diagonal, bottom left), and within chromatin state segments as annotated by ChromHMM (above the diagonal, top right) across all the autosomes (Figure 4-2). For each region the allele-specific bias

ratio defined as $\frac{\#maternal-reads - \#paternal-reads}{\#total-reads}$ was calculated. Pairwise correlations between assays were evaluated by fitting linear models to the allele-specific bias ratio data using weighted least squares. The weight for the data point for each region was calculated as the product of the total number of reads from that region from both assays. Regions with fewer than seven total reads in either assay were excluded.

For instance, we found one of the strongest allelic correlations between POL2RA and BCLAF1 binding, and one of the strongest negative correlations between H3K79me2 and H3K27me3, both at genes and chromatin state segments. For POLR2A allele-specific activity tended to be positively correlated with that of histone modifications and transcription factors, and for H3K27me3 the allelic correlations were consistently negative. Overall, we found that positive allelic correlations among the 193 ENCODE assays are stronger and more frequent than negative correlations. This may be due to preferential capture of accessible alleles and/or the specific histone modification and transcription factor, assays used in the project. Although the R^2 values for even the strongest correlations were at most 0.383, all of the pairwise correlations for which a value is indicated (i.e. not colored light gray) were statistically significant ($p < 0.05$ after multiple hypothesis correction).

4.5 Allele-specific activity is widespread across the GM12878 genome

We went on to check for allele-specific activity in a duplicate-free set of 20,518 protein-coding genes from Gencode (Harrow et al., 2012) using the measurements of allele-specific activity for gene bodies. We found 4,721 genes (including 199 on chrX) with allele-specific activity for at least one histone modification, or transcription factor, or POLR2A. The occurrence of allele-specific activity at such a high percentage of genes suggests that such activity is widespread throughout

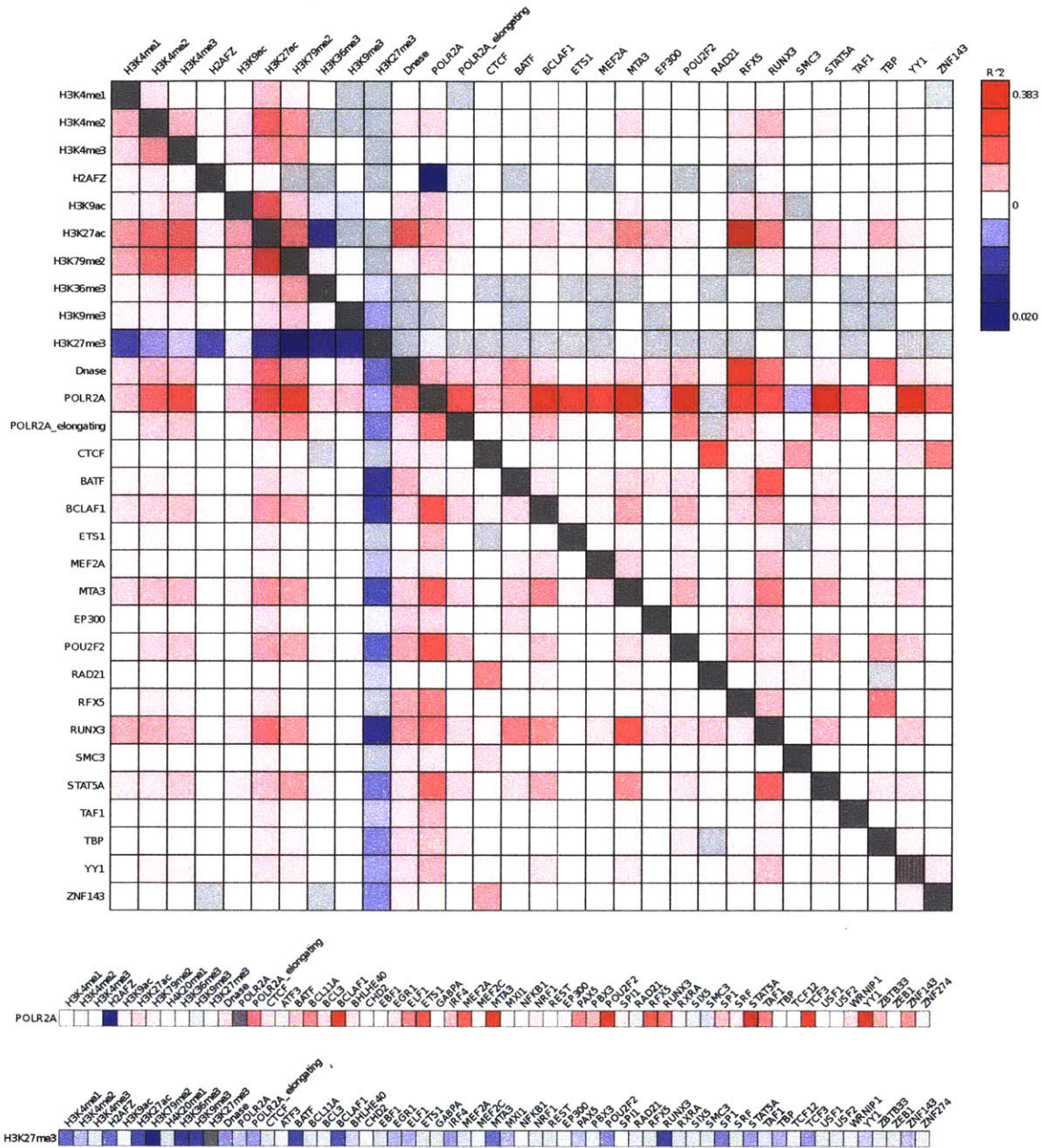


Figure 4-2: The matrix shows pair-wise correlations of allele-specific signal within single genes (below the diagonal) or within individual ChromHMM segments across the whole genome for selected DNase-seq and histone modification, transcription factor, and RNA Polymerase ChIP-seq assays. The extent of correlation is colored according to the heat-map scale indicated from positive correlation (red) through to anti-correlation (blue). Pairs of factors for which no data was available or for which correlations were not statistically significant are colored light gray. For clarity, pairwise correlations of POLR2A with other signals within ChromHMM segments, and H3K27me3 with other signals within gene bodies, are shown individually (rows below matrix). For POLR2A allele-specific activity tended to be positively correlated with that of histone modifications and transcription factors, and for H3K27me3 the allelic correlations were consistently negative.

the genome, but similar analyses would need to be performed for additional cell types and individuals to confirm this hypothesis. As shown in Figure 4-3, 158 genes had allele-specific activity for a combination of at least five transcription factors, histone modifications, or POLR2A. This suggests that it may be possible to identify sets of co-regulated genes by searching for similar patterns of allele-specific activity among groups of genes.

While it is possible that allele-specific activity for multiple transcription factors at a single gene could be due to distinct variants that explain the allele-specific binding of each factor, a simpler explanation would involve the work of pioneer transcription factors that are capable of opening up chromatin to allow for binding by other transcription factors (Zaret and Carroll, 2011). In this case, a variant disrupting a binding motif for a pioneer factor would result in chromatin being opened up on only one chromosome, resulting in allele-specific binding by all of the other transcription factors that bind to the area of open chromatin to regulate the gene.

4.6 Gaining insights into gene regulation

We examined individual genes at which we had detected allele-specific activity for POLR2A and for multiple histone modifications or transcription factors in search of examples for which allele-specific activity could provide insight into gene regulation or explain what seemed to be inconsistencies in the presence of ChIP-Seq peaks. Figure 4-4 shows RNA-Seq and ChIP-Seq data for the gene *NACC2*, which is expressed, and has ChIP-Seq peaks for POLR2A, and both H3K79me₂, a histone modification associated with transcriptional activation, and H3K27me₃, which is indicative of polycomb repression. In the absence of information about allele-specific activity this result would be rather confusing, because H3K27me₃ is associated with repression while the other signals are evidence of transcription and expression. When allele-specific activity is examined, however, we detect statistically significant allele-specific expression at two individual SNPs in the 3' UTR.

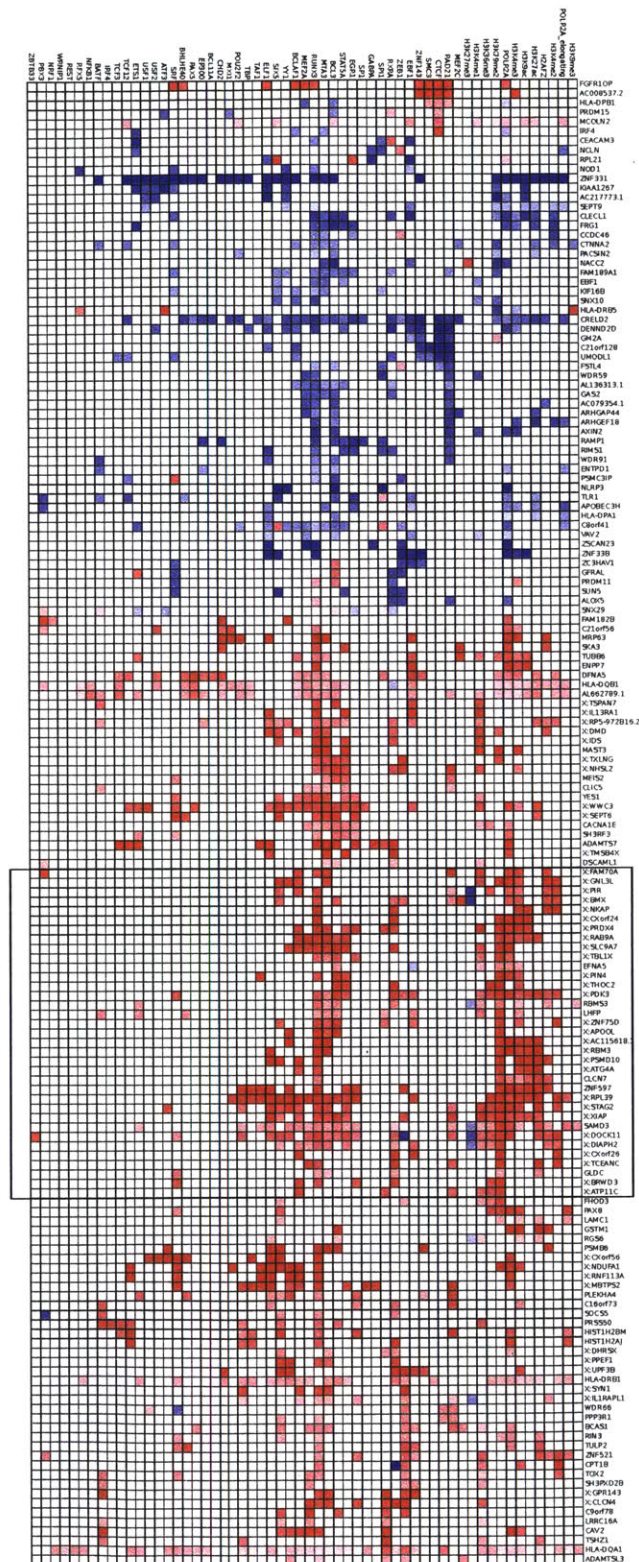


Figure 4-3: For 158 genes we detected allele-specific activity for a combination of 5 or more transcription factors, histone modifications, and/or POLR2A. Hierarchical clustering of the genes revealed that a large number of genes on the X chromosome (one copy of which is inactive) cluster together.

Allele-specific binding is also detected for POLR2A and H3K79me2 within the region spanned by the H3K79me2 ChIP-Seq peak, and for H3K27me3 within the region spanned by that ChIP-Seq peak. The allele-specific activity clearly shows that the H3K27me3 histone modification is present exclusively on the maternal chromosome, while H3K79me2, POLR2A, and expression are detected almost exclusively on the paternal chromosome.

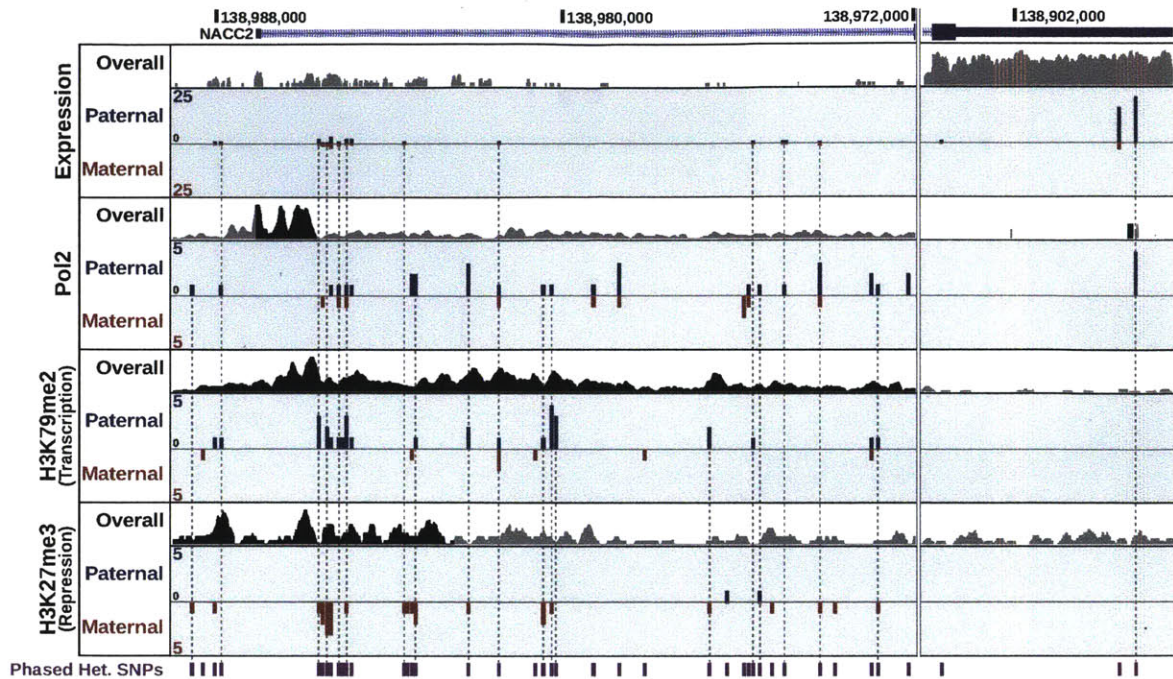


Figure 4-4: RNA-Seq (“Expression”) and ChIP-Seq data for RNA Polymerase II (“Pol2”), H3K79me2 and H3K27me3, are shown for a ≈ 16 kb region including the TSS of the gene NACC2 (left side), and for the last exon and 3’ UTR (right side). Overall signal is shown in gray, and for the ChIP-Seq signals peaks are highlighted in black. The locations of phased heterozygous SNPs are shown along the bottom of the figure. Paternal and maternal allelic read counts are indicated by bars (blue and red, respectively) below each of the overall signal tracks, and aligned with the SNPs. The gene NACC2 is expressed, as indicated by the strong signal of overall expression spanning the last exon and 3’ UTR of the gene (top right), as well as by ChIP-Seq peak calls for Pol2 overlapping the promoter and TSS, and for H3K79me2 overlapping the promoter and TSS and continuing ≈ 16 kb downstream. At the same time, a ChIP-Seq peak for H3K27me3 spanning ≈ 6 kb also overlaps the promoter and TSS. Without the context of allele-specific activity the overall signal data would be confusing, but the allele-specific activity shows that H3K27me3 is bound primarily to the maternal chromosome, while expression, Pol2 binding, and H3K79me2 binding are all primarily on the paternal chromosome.

Chapter 5

Identifying sequence variants that have functional effects

In this chapter we shift from the genome-wide view of allele-specific activity presented in the previous chapter to a more focused look at the functional effects of individual sequence variants. We show that at ChIP-Seq peaks with disrupted TF motifs allele-specific activity is correlated with the change in the PWM score of the motif. We describe a systematic approach for identifying putative causal variants in eQTLs by identifying sequence variants that alter transcription factor binding at genes where we find allele-specific binding of POLR2A or allele-specific expression.

5.1 Introduction

After examining allele-specific activity from a genome-wide perspective for the ENCODE project, we sought to explain those occurrences of allele-specific activity. Identifying variants that may be responsible for differences in transcription factor binding and gene expression is of great interest, because those differences can result in phenotypic differences including disease. Disruption of transcription factor binding motifs is thought to be one of the primary causes of these differences (Ward and Kellis, 2012b). PEGASUS is the first utility that allows insight

into the mechanisms by which sequence variations affect phenotype by detecting allelic motifs in diploid genomes and using allele-specific activity to measure their effect.

5.2 Detecting allelic motifs

When examining motif instances that overlap variants, there is most often a difference in the PWM scores of the sequence containing the reference allele and the equivalently aligned sequence containing the alternate allele. If the difference is substantial enough one of the sequences might match too poorly to be detected as an instance of the motif. Although annotations of TF motif-instances in the reference genome had been generated by the ENCODE project (Kheradpour and Kellis, 2014), novel matches to motif instances in the GM12878 genome resulting from sequence variation obviously would not have been annotated.

In order to generate a comprehensive annotation of motif instances for GM12878 we scanned both haplotypes of the GM12878 personal genome for motif matches using the same algorithms and set of motif PWMs used for the ENCODE integrative analysis. Next, we used PEGASUS to align the motif instances from the two haplotypes with each other and classify motif instances as: 1) undisrupted by variants and equivalent in both haplotypes; 2) disrupted by a variant but having equivalent PWM scores (Figure 5-1a); 3) disrupted by a variant and having an aligned match with a different PWM score (Figure 5-1b); or 4) disrupted by a variant causing such a substantial change in PWM score that the motif instance is only detected in one haplotype (Figure 5-1c). We refer to the last two types of motif instances as “allelic motifs”. We found that using only the motif instances reported for the reference genome does indeed underestimate the number of motif instances disrupted by sequence variants. Overall, we found 326,378 motif instances disrupted to such a degree that the motif instance is only called in one of the sequences, and another 70,991 motif instances that are disrupted by a variant and called as a motif instance in the sequences of both



(a) aligned motif instances with equivalent PWM scores



(b) aligned motif instances with different PWM scores



(c) motif instance only called in one haplotype

Figure 5-1: Examples of instances of an SPI1 motif overlapping variants: (5-1a) the alleles of the SNP have the same probability in the PWM; 5-1b A motif match is called in both sequences, but the motif match in the paternal sequence, with the “G” allele, has a higher PWM score; 5-1c Only the maternal sequence contains a match to the motif, because an “A” is the only base observed in occurrences of the motif at the position altered by the SNP. In the GM12878 personal genome we found 15 occurrences of motif instances that overlap variants and have equivalent PWM scores, 64 occurrences of the motif match having a higher PWM score in one haplotype than the other, and 925 occurrences where the variant disrupts the motif such that a match is only called in one haplotype.

haplotypes.

In order to detect allelic motifs PEGASUS takes as input the variants in the personal genome and the motif instances that occur in the sequences of each haplotype of the personal genome, along with the corresponding PWM match scores. Then PEGASUS checks each motif instance for overlap with variants, and attempts to pair each motif instance with the best matching motif instance from the other haplotype. In general, motif instances that overlap SNPs will align exactly to each other in both haplotypes, relative to the SNP. This is not always the case, however. For indels, the motif instances will typically be offset from each other, assuming the indel doesn’t cause the sequence of one haplotype to no longer match the motif. Even in the case of a single SNP, however, particular combinations of genome sequence and PWM structure can lead to scenarios

where, for example, the sequence containing the alternate allele does contain a match to the PWM, but it is offset by one or more bases from the motif instance containing the reference allele (and vice-versa). The occurrence of multiple motif instances in close proximity can further complicate the task of resolving how the motif instances compare with each other.

In order to facilitate the use of motif disruptions in analysis of allele specific activity and handle scenarios where motif instances are not exactly aligned, PEGASUS pairs motif instances according to several criteria. First, PEGASUS pairs motif instances that have the same PWM score and are exactly aligned with each other. Second, PEGASUS pairs offset motif instances that have the same PWM score, using a greedy algorithm to generate pairs with the smallest difference in offset relative to the common variant that they overlap. Third, PEGASUS pairs motif instances that are exactly aligned, but have different PWM scores. Finally, remaining motif instances that overlap a common variant are paired greedily according to the smallest possible difference in offset. PEGASUS outputs paired motif instances and unpaired motif instances separately from each other, and both are output along with any variants they overlap. The capability to detect allelic differences in the PWM scores of motif instances that overlap variants even when the instances are not aligned exactly with each other is a feature that is unique to PEGASUS. Providing a complete and accurate view of the motif instances in regions altered by sequence variants, greatly facilitates efforts to understand allele-specific activity.

5.3 Allele-specific activity correlates with change in motif PWM score

We next investigated whether allele-specific activity correlates with the disruption of motif instances by sequence variants. We used PEGASUS to identify ENCODE ChIP-Seq peaks which overlapped allelic motif matches for the corresponding

TFs. A window size that was the larger of the width of the peak or 1500bp was used to check for overlaps. Then, we checked for allele-specific activity using the allelic read counts that had been aggregated within the ChIP-Seq peaks, excluding peaks-motif pairs with fewer than 7 allelic reads. We used a one-sided binomial test to check whether there was allele-specific activity biased towards the haplotype for which the motif instance had a higher PWM score.

For numerous TFs we measured a strong correlation between the change in PWM score between the haplotypes and the degree of bias in allele-specific binding. For the factor SPI1, for example, we observed an R^2 of 0.82 (Pearson's) (Figure 5-2). The occurrences of allele-specific binding of SPI1 were of particular interest to study further because SPI1 plays an important role in the development of hematopoietic lineages. Also, SPI1 is a pioneer factor, capable of opening chromatin and enabling other TFs to bind, so allele-specific binding of SPI1 could lead to both allele-specific activity for other TFs and histone modifications, and to allele-specific expression.

5.4 Enrichment for allele-specific activity at GWAS loci and eQTLs

We also investigated whether allelic activity is enriched at GWAS loci and eQTLs. We checked for allelic activity occurring at SNPs from the NHGRI's GWAS Catalog (Welter et al., 2014) and in seven tissue-specific sets of eQTLs compiled by the GTEx project (Lonsdale et al., 2013; Stranger et al., 2007; Montgomery et al., 2010). Using Fisher's Exact test to check for enrichment we found that allele-specific activity is more likely to occur at GWAS and eQTL tag SNPs than at SNPs that are outside of GWAS loci and eQTLs. The enrichment was statistically significant for the GWAS loci and all of the eQTL sets (Figure 5-3). We observed the strongest effect for the lymphoblastoid-specific eQTLs (the cell type of the GM12878 cell line).

SPI1

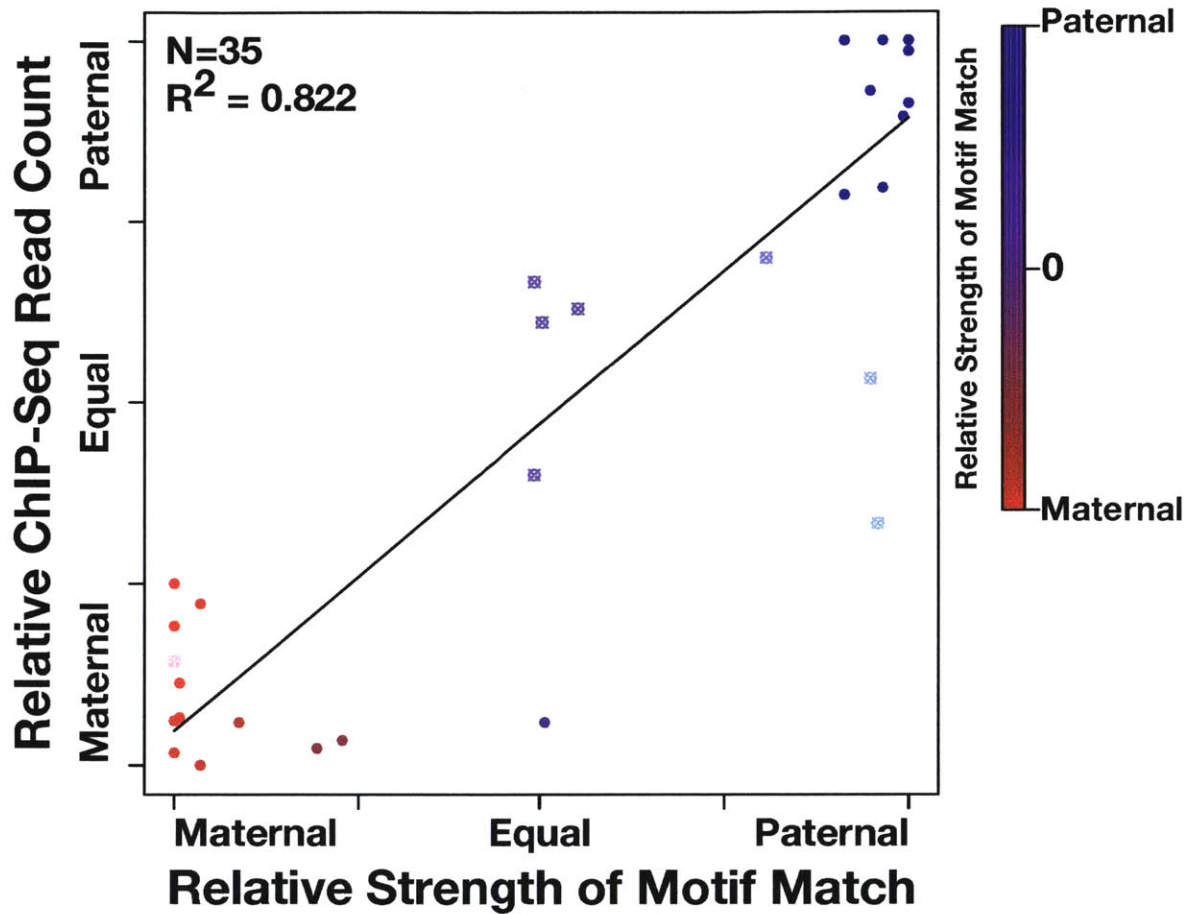


Figure 5-2: The degree of bias in allele-specific activity observed at SPI1 ChIP-Seq peaks overlapping a motif instance disrupted by a variant correlates very well, $R^2 = 0.822$, with the change in the PWM score of the motif instance. Points indicated by dots represent regions where the allele-specific activity was statistically significant and are colored according to the relative score ("strength") of the match of the motif instance to the PWM. Points indicated by circled x's and colored pink or blue represent regions where there were 7 or more allelic reads, but there was not allele-specific activity.

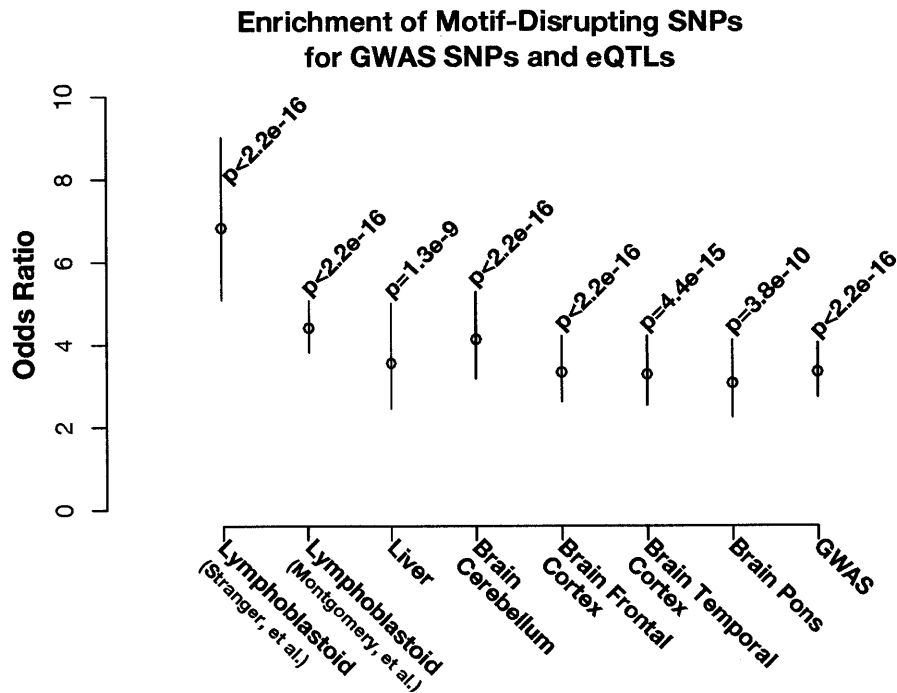


Figure 5-3: SNPs at eQTLs and GWAS Loci are three to seven times more likely to have allele-specific activity than SNPs outside those loci.

5.5 Discovering mechanisms for disease association and eQTLs

We identified 458 regions of the genome where allele-specific activity was detected at one or more ChIP-Seq peaks and at least one peak had a disruption in a motif consistent with the allele-specific activity. For each region we produced a list of genes that the region overlapped, or found the nearest gene if none were overlapped. We then narrowed down that set of regions to a set of 39 for which we detected allele-specific binding of POLR2A, or allele-specific expression at any of the genes. We ranked these regions and then sought to understand the regulation of the genes by examining the available functional data from ENCODE and checking for known eQTLs¹.

¹Although simply checking whether there were known eQTLs was straightforward, determining the direction of effect of each tag SNP and checking whether it or any SNPs in LD with it were included in the personal genome was an extremely time consuming, manual process

One of the top ranked regions was the first exon of the gene *APOBEC3H*, at which we detected coordinated allele-specific activity of four transcription factors (EBF1, EGR-1, ELF1, and RUNX3) and POLR2A, allele-specific H3K27ac histone modification, and allele-specific expression, all of which showed a bias toward the paternal chromosome (Figure 5-4). The allele-specific binding of all of the TFs was detected at ChIP-Seq peaks overlapping the first exon of the gene, where allele-specific expression was also detected. The allele-specific H3K27ac was detected across a broad peak spanning the promoter, first exon, and first intron. We found that a motif instance for EGR-1 is disrupted by a SNP, causing a change in PWM score that is consistent with the observed allele-specific activity. This region was of interest both because of the important role of EGR-1 in regulating differentiation (Yan et al., 2000), and because of the immune function of *APOBEC3H*, which creates mutations in viral genomes. In particular, changes in expression and amino acid sequence have been shown to affect its antiretroviral activity and are associated with altered resistance to HIV infection (Harari et al., 2009). Although additional experimental evidence would be required to verify that the SNP disrupting the EGR-1 motif affects the expression of *APOBEC3H*, this example demonstrates how analysis of allele-specific activity could be used to identify functional mechanisms for sequence variants associated with disease.

Two other the top ranked regions were a region around the TSS of the gene *DCAF4*, and a region approximately 25kb upstream (Figure 5-5). We detected coordinated allele-specific expression and allele-specific binding of POLR2A, and the TFs EBF1 and SPI1. Allele-specific expression and allele-specific binding of POLR2A were both detected in the first exon of the gene. The allele-specific binding of EBF1 was detected at a ChIP-Seq peak in the first intron of the gene, only about 1kb downstream of the TSS. The ChIP-Seq data for this peak is actually the data shown in Figure /reffig:pgc:ref-bias-at-ebf1-peak. The bias in EBF1 binding primarily to the maternal haplotype was consistent with the disruption of an EBF1 motif directly under the center of the peak resulting in a much stronger match to the PWM on the maternal chromosome. At the region 25kb upstream

we detected allele-specific binding of the transcription factor SPI1 on the maternal chromosome. This, too, was consistent with the disruption of an SPI1 motif under the ChIP-Seq peak. The SNP alters a position that shows high specificity for an “A” and a motif match was detected only on the maternal chromosome.

Both SPI1 and EBF1 are known to play important roles in the development of lymphoblastoid lineages (Kee and Murre, 2001). In order to find more direct evidence that the SNPs causing allele-specific binding of the TFs were also responsible for the allele-specific expression, we checked for known eQTLs for DCAF4 in the GTEx datasets. We found that both regions had been identified as eQTLs, but that no causal variants had been proposed (Pickrell et al., 2010; Zeller et al., 2010). For the eQTL at the EBF1 ChIP-Seq peak, the SNP most highly associated with change in expression of DCAF4 (rs1076458) is the very same SNP disrupting the EBF1 motif. The SNP disrupting the SPI1 motif is in LD ($D' = 1$) with the most highly associated SNP from that eQTL (rs7144189) (Ward and Kellis, 2012a). This strongly suggests that the causal variants for those eQTLs are the SNPs that disrupt the TF binding motifs and the mechanism by which they alter expression is by altering the binding of the TFs.

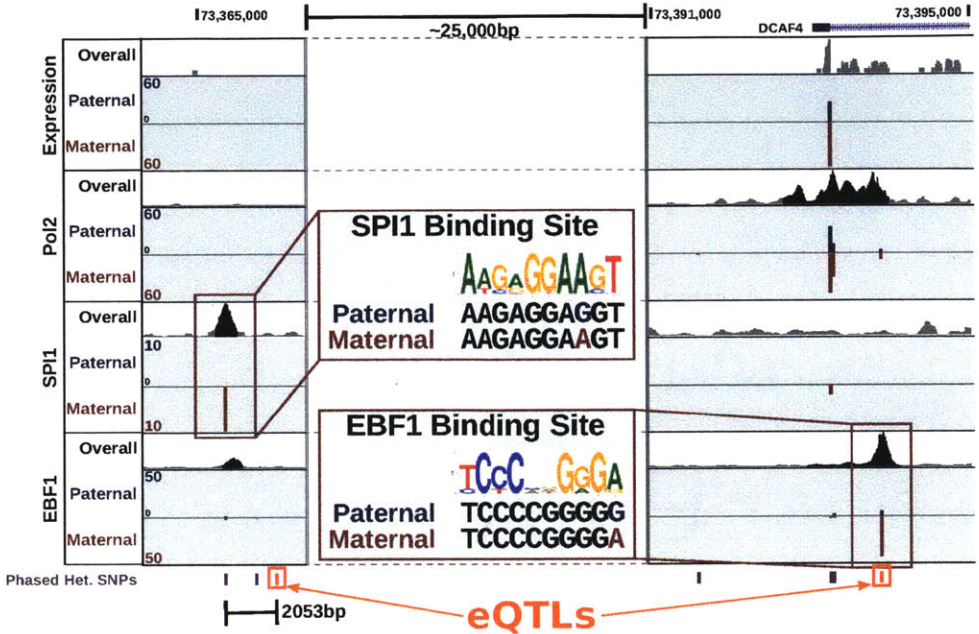


Figure 5-5: Allele-specific expression and allele-specific binding of RNA Polymerase II (“Pol2”), both biased towards the maternal chromosome, are detected at the first exon of the gene DCAF4 (upper right). Allele-specific binding to the maternal chromosome is also detected for the TFs EBF1 (bottom right) and SPI1 (left) at ChIP-Seq peaks located approximately 1kb downstream of the TSS, and about 25kb upstream of the TSS, respectively. The maternal bias in allele-specific binding of both TFs is consistent with the changes in PWM scores caused by SNPs disrupting motifs under both peaks. In the case of the EBF1 motif instance, the alternate allele, “A”, produces a better match to the PWM on the maternal chromosome. Similarly, an SPI1 motif match is only detected on the maternal chromosome, because the alternate allele, “G”, on the paternal chromosome occurs at a motif position with a high specificity for “A”. Both EBF1 and SPI1 are known to play important roles in the development of lymphoblastoid lineages. We also found that the SNP altering the EBF1 motif has been reported as an eQTL for DCAF4, along with a SNP 2kb downstream of the SPI1 motif (orange, bottom row).

Chapter 6

Conclusion

6.1 Summary of results

This thesis presented methods for creating personal genomes in order to improve the analysis of data from sequencing-based assays, and for measuring allele-specific activity to gain insights into gene regulation and the functional effects of sequence variants.

We began by presenting a method for creating personal, diploid genomes (Chapter 2) that applies a novel formulation of a Hidden Markov Model to produce maximum-likelihood assignments of variants to haplotypes. We implemented a modified version of the Viterbi decoding algorithm to resolve inconsistencies in variant calls and produce accurate assignments efficiently. The method for assigning variants to haplotypes and creating personal genomes is the core of a software package of Personal Genome and Allele-Specific Utilities (PEGASUS). We compared personal genomes created by PEGASUS and AlleleSeq and showed that PEGASUS does a better job of generating valid haplotype assignments for overlapping variant calls.

In Chapter 3 we described how other utilities included in PEGASUS facilitate the incorporation of sequenced reads aligned to personal genomes into workflows designed for reads aligned to a reference genome. We also presented a variant-aware method for detecting PCR duplicates and described the method

used by PEGASUS to accurately measure allele-specific activity. We compared the measurements of allele-specific activity made by PEGASUS and AlleleSeq and showed the differences resulting from the quality score requirements implemented by PEGASUS. Finally, we described how we improved our sensitivity to detect allele-specific activity by taking advantage of phasing information and measuring allele-specific activity at functional elements rather than just individual variants.

In the final two chapters we presented the results of a study characterizing allele-specific activity at a genome-wide level and an analysis focused on explaining the effects of individual sequence variants. The first study, conducted as part of the ENCODE Project Consortium's integrative analysis, showed that allele-specific activity was widespread throughout the GM12878 genome and detected allelic correlations at a genome-wide level. We also provided an example of how allele-specific activity could be used to explain conflicting signals of gene regulation. In the analysis presented in Chapter 5 we first examined the correlations between allele-specific activity and the disruptions in motif instances caused by sequence variants. Next, we showed that by using allele-specific activity to identify sequence variants that alter transcription factor binding we could identify putative causal variants in eQTLs as well as variants that may affect disease phenotype.

6.2 Future work

The analysis presented in this thesis was limited in that it was applied primarily to a single lymphoblastoid cell line. As the generation of whole genome sequences and high quality variant calls becomes more commonplace there will be more opportunities to make use of measurements of allele-specific activity to help explain the functional effects of sequence variants. Already, we are in the process of conducting an analysis of allele-specific activity involving more than a dozen cell lines on which the Roadmap Epigenomics Mapping Consortium performed a

variety of functional assays.

By examining allele-specific activity data for multiple cell lines and cell types it will be possible to verify the predictions of functional effects based on analysis of allele-specific activity in single cell lines, distinguish cis-effects and trans-effects, and improve predictions of the effects of sequence variants. Additional, targeted, functional assays performed on a small number of cell lines or samples selected for having appropriate genotypes could be used to verify predictions. Massively parallel reporter assays could also be used to test predictions of effects on gene expression. Cis-effects should be consistent for difference cell lines with the same genotype, and trans-effects should be detectable as effects that remain consistent despite cell lines having different genotypes at the location where the effect is measured. The ability to distinguish these effects will both help provide additional insight into the functional effects of sequence variants and improve predictions of the effects of sequence variants.

Whereas eQTL and GWA studies require hundreds or thousands of individuals and it is extremely challenging to identify causal variants in the loci that are associated with the trait or disease, allele-specific activity measured in a single or small number of cell lines can provide very direct evidence for the functional effects of sequence variants. Analyses of allele-specific activity have the potential to both help explain the associations detected by eQTL and GWA studies and serve as an alternative approach to identify variants that may affect gene expression and disease.

Bibliography

- Altshuler, D., Daly, M. J., and Lander, E. S., 2008. Genetic mapping in human disease. *science*, **322**(5903):881–888.
- Bengio, Y. and Frasconi, P., 1995. An input output hmm architecture. *Advances in neural information processing systems*, :427–434.
- Carrel, L. and Willard, H. F., 2005. X-inactivation profile reveals extensive variability in x-linked gene expression in females. *Nature*, **434**(7031):400–404.
- Consortium, . G. P. et al., 2012a. An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**(7422):56–65.
- Consortium, E. P. et al., 2012b. An integrated encyclopedia of dna elements in the human genome. *Nature*, **489**(7414):57–74.
- Consortium, T. E. P., 2011. A user’s guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology*, **9**(4):e1001046.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., *et al.*, 2011. The variant call format and vcftools. *Bioinformatics*, **27**(15):2156–2158.
- Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., and Pritchard, J. K., 2009. Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data. *Bioinformatics*, **25**(24):3207–3212.
- DeVeale, B., Van Der Kooy, D., and Babak, T., 2012. Critical evaluation of imprinted gene expression by rna-seq: a new perspective. *PLoS Genet*, **8**(3):e1002600.
- Dondorp, W. J. and de Wert, G. M., 2013. The “thousand-dollar genome”: an ethical exploration. *European Journal of Human Genetics*, **21**:S6–S26.
- Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G., 1998. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.
- Ernst, J. and Kellis, M., 2012. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, **9**(3):215–216.

- Garbis, S., Lubec, G., and Fountoulakis, M., 2005. Limitations of current proteomics technologies. *Journal of Chromatography A*, **1077**(1):1–18.
- Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., *et al.*, 2003. The international hapmap project. *Nature*, **426**(6968):789–796.
- Gossett, A. J. and Lieb, J. D., 2010. DNA immunoprecipitation (DIP) for the determination of DNA-Binding specificity. *Cold Spring Harbor Protocols*, **2008**(3):pdb.prot4972.
- Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G. P., Haig, D., and Dulac, C., 2010. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *science*, **329**(5992):643–648.
- Guo, Y., Papachristoudis, G., Altshuler, R. C., Gerber, G. K., Jaakkola, T. S., Gifford, D. K., and Mahony, S., 2010. Discovering homotypic binding events at high spatial resolution. *Bioinformatics*, **26**(24):3028–3034.
- Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., *et al.*, 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature biotechnology*, **28**(5):503–510.
- Harari, A., Ooms, M., Mulder, L. C., and Simon, V., 2009. Polymorphisms and splice variants influence the antiretroviral activity of human apobec3h. *Journal of virology*, **83**(1):295–303.
- Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., *et al.*, 2012. Gencode: the reference human genome annotation for the encode project. *Genome research*, **22**(9):1760–1774.
- Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., *et al.*, 2006. The ucsc genome browser database: update 2006. *Nucleic acids research*, **34**(suppl 1):D590–D598.
- Hirschhorn, J. N. and Daly, M. J., 2005. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, **6**(2):95–108.
- Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., *et al.*, 2012. Integrative annotation of chromatin elements from encode data. *Nucleic acids research*, :gks1284.
- Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K., 2008. Genome-wide identification of in vivo protein–dna binding sites from chip-seq data. *Nucleic acids research*, **36**(16):5221–5231.

- Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., *et al.*, 2010. Variation in transcription factor binding among humans. *Science*, **328**(5975):232–235.
- Kee, B. L. and Murre, C., 2001. Transcription factor regulation of b lineage commitment. *Current opinion in immunology*, **13**(2):180–185.
- Kheradpour, P. and Kellis, M., 2014. Systematic discovery and characterization of regulatory motifs in encode tf binding experiments. *Nucleic acids research*, **42**(5):2976–2987.
- Kidd, J. M., Cheng, Z., Graves, T., Fulton, B., Wilson, R. K., and Eichler, E. E., 2008. Haplotype sorting using human fosmid clone end-sequence pairs. *Genome research*, **18**(12):2016–2023.
- Kouzarides, T., 2007. Chromatin modifications and their function. *Cell*, **128**(4):693–705.
- Kuhn, R. M., Haussler, D., and Kent, W. J., 2012. The ucsc genome browser and associated tools. *Briefings in bioinformatics*, :bbs038.
- Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., *et al.*, 2015. Integrative analysis of 111 reference human epigenomes. *Nature*, **518**(7539):317–330.
- Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C., Beecham, G. W., *et al.*, 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer’s disease. *Nature genetics*, **45**(12):1452–1458.
- Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L., *et al.*, 2009. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biol*, **10**(3):R25.
- Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**(14):1754–1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., *et al.*, *et al.*, 2009. The sequence alignment/map format and samtools. *Bioinformatics*, **25**(16):2078–2079.
- Li, H., Ruan, J., and Durbin, R., 2008a. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome research*, **18**(11):1851–1858.
- Li, R., Li, Y., Kristiansen, K., and Wang, J., 2008b. Soap: short oligonucleotide alignment program. *Bioinformatics*, **24**(5):713–714.

- Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., *et al.*, 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**(7370):476–482.
- Liu, C.-M., Wong, T., Wu, E., Luo, R., Yiu, S.-M., Li, Y., Wang, B., Yu, C., Chu, X., Zhao, K., *et al.*, 2012. Soap3: ultra-fast gpu-based parallel alignment tool for short reads. *Bioinformatics*, **28**(6):878–879.
- Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.*, 2013. The genotype-tissue expression (gtex) project. *Nature genetics*, **45**(6):580–585.
- Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., *et al.*, 2003. TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, **31**(1):374–378.
- McDaniell, R., Lee, B.-K., Song, L., Liu, Z., Boyle, A. P., Erdos, M. R., Scott, L. J., Morken, M. A., Kucera, K. S., Battenhouse, A., *et al.*, 2010. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, **328**(5975):235–239.
- Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. T., 2010. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, **464**(7289):773–777.
- Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., *et al.*, 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, **44**(9):981.
- Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S., 2011. Genotype and snp calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**(6):443–451.
- of the Psychiatric Genomics Consortium, S. W. G. *et al.*, 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**(7510):421–427.
- Pandey, R. V. and Schlötterer, C., 2013. Distmap: a toolkit for distributed short read mapping on a hadoop cluster. *PloS one*, **8**(8):e72614.
- Pearson, W. R. and Lipman, D. J., 1988. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, **85**(8):2444–2448.
- Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K., *et al.*, 2010.

- Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, **464**(7289):768–772.
- Rivas-Astroza, M., Xie, D., Cao, X., and Zhong, S., 2011. Mapping personal functional data to personal genomes. *Bioinformatics*, **27**(24):3427–3429.
- Rockman, M. V. and Kruglyak, L., 2006. Genetics of global gene expression. *Nature Reviews Genetics*, **7**(11):862–872.
- Rosenbloom, K. R., Dreszer, T. R., Pheasant, M., Barber, G. P., Meyer, L. R., Pohl, A., Raney, B. J., Wang, T., Hinrichs, A. S., Zweig, A. S., *et al.*, 2010. Encode whole-genome data in the ucsc genome browser. *Nucleic acids research*, **38**(suppl 1):D620–D625.
- Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., *et al.*, 2011. Alleleseq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology*, **7**(1):522.
- Sanger, F., Nicklen, S., and Coulson, A. R., 1977. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, **74**(12):5463–5467.
- Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F., Barbalic, M., Gieger, C., *et al.*, 2011. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*, **43**(4):333–338.
- Seto, E., Shi, Y., and Shenk, T., 1991. YY1 is an initiator sequence-binding protein that directs and activates transcription in vitro. *Nature*, **354**(6350):241–245.
- Stevenson, K. R., Coolon, J. D., and Wittkopp, P. J., 2013. Sources of bias in measures of allele-specific expression derived from rna-seq data aligned to a single reference genome. *BMC genomics*, **14**(1):536.
- Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., *et al.*, 2007. Population genomics of human gene expression. *Nature genetics*, **39**(10):1217–1224.
- Trapnell, C., Pachter, L., and Salzberg, S. L., 2009. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, **25**(9):1105–1111.
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L., *et al.*, 2012. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols*, **7**(3):562–578.

- Turro, E., Su, S.-Y., Gonçalves, Â., Coin, L., Richardson, S., Lewin, A., et al., 2011. Haplotype and isoform specific expression estimation using multi-mapping rna-seq reads. *Genome Biol*, **12**(2):R13.
- van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. K., 2015. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods*, **12**(11):1061–1063.
- van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C., 2014. Ten years of next-generation sequencing technology. *Trends in genetics*, **30**(9):418–426.
- Wang, Z., Gerstein, M., and Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**(1):57–63.
- Ward, L. D. and Kellis, M., 2012a. Haploreg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*, **40**(D1):D930–D934.
- Ward, L. D. and Kellis, M., 2012b. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*, **30**(11):1095–1106.
- Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al., 2014. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, **42**(D1):D1001–D1006.
- Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., et al., 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, **46**(11):1173–1186.
- Yan, S.-F., Fujita, T., Lu, J., Okada, K., Zou, Y. S., Mackman, N., Pinsky, D. J., and Stern, D. M., 2000. Egr-1, a master switch coordinating upregulation of divergent gene families underlying ischemic stress. *Nature medicine*, **6**(12):1355–1361.
- Yoon, B.-J. and Vaidyanathan, R., 2004. Rna secondary structure prediction using context-sensitive hidden markov models. In *Biomedical Circuits and Systems, 2004 IEEE International Workshop on*, pages S2–7. IEEE.
- Zaret, K. S. and Carroll, J. S., 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes & Development*, **25**(21):2227–2241.
- Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K., Rossmann, H., et al., 2010. Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PloS one*, **5**(5):e10693–e10693.