Evaluating the Efficacy of Epigenetic Imputation in CD4$^+$ Regulatory T Cells

Kiran Kumar Thurimella

Wellcome Sanger Institute

University of Cambridge
Darwin College

This dissertation is submitted for the degree of Master of Philosophy in Biological Sciences
August, 2018

Statement of Length: This dissertation does not exceed the word limit set forth by the Degree Committee of Biological Science of 20,000 words. The count for this thesis is 10,386 words.

Abstract

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) for histone marks has been widely used to characterize non-coding elements of the genome which control gene expression and can contribute to disease (Trynka et al. 2013). Although, ChIP-seq is a well established method, depending on the mark and both availability and the type of cellular material the protocol can be challenging and sensitive to technical variability (Park 2009). This can result in inaccurate read coverage or low sequencing depth. Methods such as epigenetic imputation (Zhou & Troyanskaya 2015; Alipanahi et al. 2015; Ernst & Kellis 2015; Bock & Lengauer 2008), which statistically infer missing or unobserved regions of the non-coding genome, can be used to potentially improve the overall quality of the data. In this study, I evaluated the software tool ChromImpute using public and internal data from various T cell populations to evaluate the performance of imputation.

Firstly, I tested ChromImpute using data from different T cell populations, including $CD4^+$ Effector Memory, $CD4^+$ Central Memory, $CD4^+$ Regulatory T cells, $CD3^+$ Thymocyte, $CD3^-$ Thymocyte, $CD4^+$ Alpha Beta T cells, generated as a part of the BLUEPRINT consortium (Adams et al. 2012). For these samples, I imputed five chromatin marks: H3K27ac, H3K4me1, H3K4me3, H3K9me3 and H3K27me3 using the ROADMAP (Bernstein et al. 2010) and ENCODE (ENCODE Project Consortium 2012; ENCODE Project Consortium 2004) reference data. Next, I applied ChromImpute to data from three regulatory T cell (Treg) samples generated in our lab, using a combination of BLUEPRINT and in-house data as a reference compendium. To evaluate the imputation performance, I focused on the H3K27ac and H3K4me1 marks, as these marks had the greatest sequencing depth. Finally, I imputed data for an additional 11 Treg samples to assess if ChromImpute is able to preserve genotypically driven variability.

This study provided insights into the performance of ChromImpute for histone ChIP-seq data. My results indicate that ChromImpute preserves global structure of chromatin while reducing noise, filling in missing data and correcting for experimental biases. ChromImpute also reduces the impact of technical variability in ChIP-seq data. However, I observe that imputation does not capture the genotypic variability.

Table of Contents

1. Introduction

The immune system comprises of multiple molecules, cells and organs, working together to protect the host from harmful pathogens (Akira et al. 2006). The recognition of foreign and harmful organisms by the immune system is not straightforward, and there are multiple mechanisms involved in preventing immune activation against healthy cells or harmless foreign antigens, like food (Wang et al. 2014). Autoimmune diseases arise from a malfunction along this intricate network of interactions (Rioux & Abbas 2005) and are a grappling problem due to their debilitating nature and large treatment costs (Rosenblum et al. 2015). Cytokine antagonists, such as TNF-α, have shown great efficacy in remediating common autoimmune diseases (Kriegler et al. 1988; Colombel et al. 2007; Rutgeerts et al. 2005; Furst 2010). However, treatment is often received after the disease is in a late stage of pathogenesis and to improve current therapeutic alternatives it will be important to understand what triggers the initial development of disease.

Some autoimmune diseases are characterized by an imbalance between effector T cells and regulatory T cells (Tregs) (Bluestone et al. 2008). Tregs play a key role in immune homeostasis as major suppressors of activated T cells and non-functional Tregs are thought of as a major player in the downstream cascade of autoimmune disease (Kronenberg & Rudensky 2005; Sakaguchi et al. 2008; Li & Zheng 2015). Tregs control and regulate T cell activation which in turn helps to maintain self-tolerance (Sakaguchi et al. 1995). As such, understanding the role of genetics in the development of Tregs is critical in unravelling the reason for Treg malfunction and autoimmunity. Many studies have begun to investigate the influence of genetics on Treg dysfunction (Guerini et al. 2012; Pidasheva et al. 2011; Raychaudhuri et al. 2012; Zhou et al. 2011) and have identified certain genetic polymorphisms in genes such as *CTLA-4, STAT3, FOXP3,* and *IL-2,* which may affect the development and function of Tregs (Fletcher et al. 2009; Encinas et al. 1999; Holland et al. 2007; Atabani et al. 2005).These findings have also implicated the polygenic nature of

autoimmune diseases, which has motivated the expansion of Genome-wide association studies (GWAS) have discovered that many disease associated single nucleotide polymorphisms (SNPs) fall in non-coding regions of the genome (Wellcome Trust Case Control Consortium et al. 2010). These regions often carry epigenetic marks and can play a role in regulating gene expression (Trynka et al. 2013; Trynka & Raychaudhuri 2013).

Given these findings, to understand disease it will be crucial to characterize the epigenome of immune cell types. Specifically, understanding how polymorphisms affect the function of Tregs could help elucidate the mechanisms behind inflammatory cascades in autoimmune disease. However, Tregs are a rare cell type and profiling them can be expensive, labor intensive and error prone (Pierini et al. 2014). One major issue is the scarcity of sequencing data from rare cell populations such as Tregs (Librado & Rozas 2009). Genotype imputation methods have been developed to help boost power, and recover missing data in various GWAS (Y. Li et al. 2009). However, only recently have methods to analyze and impute epigenetic information been developed (Ernst & Kellis 2015; Zhou & Troyanskaya 2015). These methods can be used to impute data for rare immune cell types, thus capturing better epigenetic profiles. Applying this approach to Tregs would help us build a foundation for functionally characterizing this cell type and its role in autoimmune disease.

In this chapter, I will summarize the current understanding of GWAS and give an introduction to imputation and epigenomics in Treg cells. I will then discuss some of the limitations of assaying various epigenetic marks and show how these may be resolved using imputation methods.

1.1 Challenges in studying complex diseases

1.1.1 The genetic architecture of common diseases

Genome-wide association studies (GWAS) are a type of study designed to determine which genetic variants are associated with a given trait (for example, a disease) and have been widely used to study common diseases. GWAS works by comparing the genotypes of individuals affected by the disease with the genotypes of a control group of individuals and are based on the concept of linkage disequilibrium (LD), the association between various alleles at different loci in a given individual (which is determined by ethnicity) (Bush & Moore 2012). Generally speaking, loci that are close together in the genome have a strong LD. Since these alleles are present on these haplotypes they tend to be inherited more often together, which leads to these alleles being correlated together. A haplotype is a group of alleles that are inherited in unison from one parent (Gabriel et al. 2002). Based on this concept, one can select a fraction of common variants in a population and rank them based on a p-value, which is a function of the number of independent variant tests performed (Schork et al. 2013). Next, rare variants can be inferred from already sequenced fragments and be imputed based on the genotyped common variants. This methodology provides a framework for mapping variants associated to the disease. When performing GWAS, individuals are mostly genotyped based on their SNPs, as SNPs are the most common form of genetic variation (Martin et al. 2000). Because of the large number of genetic variants in the genome and the relatively small contribution of each of them to disease, GWAS studies generally need large cohort sizes; the larger the sample size, the more accurately one can detect small effect sizes i.e. the statistical power increases (Spencer et al. 2009).

During the last decade, GWAS have identified thousands of associations between diseases and SNPs (Visscher et al. 2017). These studies have revealed that often diseases are caused by the added effect of multiple alleles spread across the genome (complex traits) (Raychaudhuri 2011) and that such phenotypes are often polygenic (Das et al. 2006). Additionally, most common complex diseases also have a significant environmental component (Ramos & Olden 2008).

1.1.2 Interpretation of disease associated variants is challenging

Though GWAS studies have revealed many different disease associations, there are several challenges that remain. The majority of variants identified in these studies are unlikely to be disease causative (Anderson et al. 2011). This is partly because of the low power and resolution GWAS studies have for variants of small effect sizes (Manolio et al. 2009). This makes it difficult to translate results from GWAS studies into detailed functional mechanisms underlying disease (Edwards et al. 2013). Even though a number of etiologies have been uncovered through associated genes and pathways, there remains a wide gap between these studies and clinically relevant associations (Manolio 2013; Varmus 2010; Evans et al. 2011).

Moreover, many disease associated SNPs fall in intergenic and intronic, non-coding regions of the genome (Blattler et al. 2014; Freedman et al. 2011; Tehranchi et al. 2016). As such, functional follow up on these regions is challenging. There is no knowledge of which gene these non-coding variants affect or by which mechanism they affect it. Moreover, many of these variants have effects which can be cell-type specific (Korte & Farlow 2013). However, often we do not know the gene expression in specific cell populations that may be directly affected by the variants. Thus, it has become crucial to go beyond associated loci and investigate what is mechanistically occurring in these regions and in which cell type those effects are observed. All of these hurdles require careful consideration and will be vital in ultimately using human genetics to drive translational medicine.

1.2 Epigenetics can be used to study non-coding regions

1.2.1 Gene expression and cell function are regulated at the epigenetic level

In the eukaryotic cell nucleus, chromatin is made of packed DNA that is wrapped around an octamer of histones known as nucleosome (Lorch et al. 2010). Gene expression is dependent on how densely packed chromatin is, a characteristic called chromatin accessibility (Tsompana & Buck 2014). The positioning of nucleosomes in a genome affects the accessibility of the transcriptional machinery to elements such

as transcription factor binding sites at gene promoters and enhancers (Radman-Livaja & Rando 2010). In addition, the binding of transcription factors themselves affect the accessibility of chromatin (Thurman et al. 2012). Changes in chromatin accessibility influence the function of a cell through regulation of gene expression (Shu et al. 2011; Korber et al. 2004; Schones et al. 2008).

Within a histone complex, there are four basic histones that form the octamer: H2A, H2B, H3 and H4 (Karlić et al. 2010). These histones can undergo post-translational modifications, which are associated with changes in chromatin accessibility and regulatory function (Bannister & Kouzarides 2011; Kouzarides 2007; Rea et al. 2000). These modifications can be categorized as repressive methylations or activating methylations/acetylations **(Figure 1.1)** (Barski et al. 2007; Liang et al. 2004; Benevolenskaya 2007; Rosenfeld et al. 2009).

Histone Modification	Mark Region	Gene Expression Status
H3K27ac	Proximal/Distal to TSS	activation
H3K4me1	Proximal/Distal to TSS	activation
H3K4me3	Near promotors	activation
H3K36me3	Distal to TSS	repression
H3K27me3	Enriched throughout TSS	repression
H3K9me3	Located at gene bodies	repression
H3K9ac	Proximal to TSS	activation

Figure 1.1 Histone modifications used in epigenetic imputation project This table represents the different histone modifications that were tested in the Tier 1 ChromImpute (Ernst & Kellis 2015) study. These histone modifications have diverse effects on transcription and are well documented in several experiments (Barski et al. 2007; Liang et al. 2004; Benevolenskaya 2007; Rosenfeld et al. 2009; Gates et al. 2017; Lauberth et al. 2013; Liu et al. 2015).

1.2.2 Chromatin profiling can be used to functionally annotate non-coding regions

Regulatory chromatin state can be inferred by profiling the histone marks or transcription factors occupying the chromatin, using chromatin immunoprecipitation followed by sequencing (ChIP-seq or ChIPmentation) (Schmidl et al. 2015; O'Geen et al. 2011). Another assay developed for profiling open chromatin regions is the transposase-accessible chromatin followed by sequencing (ATAC-seq) (Buenrostro et al. 2013). ChIP-seq against histone marks assays regions tagged by a specific chromatin modification, first histone proteins are cross-linked with the DNA bound to them. Following the cross-linking, the cells are lysed (cell membranes are broken) (Landt et al. 2012). Next sonication is used to shear the DNA in the region (O'Geen et al. 2011). Then, chromatin is immunoprecipitated using antibodies against the chromatin mark of interest, proteins are removed and DNA is sequenced (Landt et al. 2012). ChIPmentation follows the same process as ChIP-seq but combines it with sequencing library preparation done by Tn5 transposase (Schmidl et al. 2015).

ATAC-seq uses the Tn5 transposase to randomly insert sequencing adaptors in accessible regions of a chromatin sample. The Tn5 acts on DNA located within regions of open chromatin. The resulting cutting sites have specific Tn5 adapters that can then be amplified via PCR (Buenrostro et al. 2013) and ultimately sequenced.

Over 90% of disease associated GWAS variants are non-coding and ChIP-seq is a proficient method to begin to annotate non-coding regions (ENCODE Project Consortium et al. 2007; Zhang & Lupski 2015; Robertson et al. 2007; Mikkelsen et al. 2007; Valouev et al. 2008; Hrdlickova et al. 2014). To drive towards the mechanistic cause of disease, GWAS and ChIP-seq can be combined to find which SNPs are more likely to be active and also in a given cell type. ChIP-seq has been able to identify enhancers and different active elements in specific tissues (Visel et al. 2009). Different tools, such as the GREAT method (McLean et al. 2010) or the

JASPAR database (Portales-Casamar et al. 2010), have been built on the foundation of ChIP-seq to find transcription factor binding sites.

1.2.3 High-throughput chromatin profiling has technical limitations

Often times, running chromatin assays can be time intensive and costly. For example, despite several improvements since its first use it in 2007, ChIP-seq is still biased towards GC rich fragments during library preparation, selection and sequence amplification (Quail et al. 2008). And whether a library is prepared for paired-end or single-end sequencing can also influence the quality of the reads generated (Chen et al. 2012), which can add to the financial burden when designing a ChIP-seq experiment. Moreover, when there is an inadequately small number of reads, enriched regions became ill defined (Park 2009; Zhang & Pugh 2011). This causes the minimum viable sequencing depth for any human samples to be around 40-50 million reads, which can be costly (Jung et al. 2014).

Given the previous limitations, designing and conducting ChIP-seq assays can be expensive and time consuming. Consortia such as ROADMAP (Bernstein et al. 2010), ENCODE (ENCODE Project Consortium 2004; ENCODE Project Consortium 2012) and BLUEPRINT (Adams et al. 2012) have generated large amounts of epigenetic data across many different cell types. These datasets have shown that chromatin marks tend to correlate with each other, which makes a strong case for using imputation to aid in epigenetic profiling using chromatin assays, as described in the next section (Stunnenberg et al. 2016; ENCODE Project Consortium 2012).

1.3 Imputation can be used to improve epigenetic profiling

1.3.1 Imputation is routinely used to improve genotyping quality

High-throughput technologies such as genotyping or epigenetic profiling often exhibit technical biases. One way to minimise their impact is by applying imputation. For instance, genotype imputation has been used to augment GWAS studies, boost power, fine-map associations and help draw conclusions across various GWAS (Y.

Li et al. 2009). Fine-mapping is the process of associating causality to disease variants with standardized probabilities (Spain & Barrett 2015). Imputation does so by predicting genotypes that are not directly assayed in different samples (Marchini & Howie 2010). Providing genotype information for every individual base pair in the genome is infeasible due to expense and time cost (Mertes et al. 2011). Imputation can help address these issues by taking a SNP microarray that surveys a subset of different SNPs and predicts the genotypes (Howie et al. 2009). Genotype imputation is built off of common haplotype structure, where a reference of these haplotypes is used to predict missing genotype values in a set of individuals where a set of genotyped SNPs exists (Browning & Browning 2007). After generating this set of imputed SNPs, the number of associations can be tested with a larger set of SNPs. With increased power, the causal variants in the set of SNPs can be targeted further for analysis (NCI-NHGRI Working Group on Replication in Association Studies et al. 2007).

One common issue that has limited GWAS studies is the diversity in both genotyped samples and haplotype references (Hunter & Kraft 2007). However, the increasing diversity of HapMap 3 consortium (The International HapMap 3 Consortium et al. 2010) has helped further improve imputation references (Trynka et al. 2013). Due to these advancements, imputation has become a standard practice when performing GWAS, as sequencing and genotyping can themselves be expensive and prone to errors (Visscher et al. 2017). This method has provided pivotal planning for scientists to conduct experiments by genotyping a selection of variants in a sample and imputing the remainder ones (Roshyara et al. 2016). It is worth noting that many different protocols which capture functional information such as microarrays, RNA-seq, ChIP-seq and, ATAC-seq, also contain missing values (Troyanskaya et al. 2001) and can thus benefit from imputation.

A typical genotype imputation model begins with separating the data into: 1) typed in the reference and in the study, 2) untyped in the study but typed in the reference.

The SNPs typed in the study are next matched to see if there is a reference haplotype that best fits them (Troyanskaya et al. 2001). It is then assumed that the SNPs untyped in the study follow the same structure as the reference and hence the SNPs are imputed (Y. Li et al. 2009; Marchini & Howie 2010; Troyanskaya et al. 2001).

A substantial proportion of imputation methods are based on Hidden Markov Models (HMMs). HMMs assume the existence of a series of events, where the probability of any given event depends only on the state of the previous one (Bini et al. 2005). The adjective “hidden” refers to the idea that the state of an event can be unobserved (i.e. hidden). HMMs have been applied to a variety of problems in biology, for example DNA motif prediction (Eddy 1998; Krogh 1998) and genotype imputation. Genotyping methods based on hidden Markov models generally estimate the haplotype (phasing) of each individual from the reference panel only (Howie et al. 2012). One pitfall to this approach is that it does not take into account the study size when trying to provide missing information (Spencer et al. 2009). Other imputation algorithms use machine learning principles to learn from a set of features and more accurately predict missing values (Jerez et al. 2010). The most robust imputation methods rely on a combination of these two approaches (Cantor et al. 2010).

It is important to verify if the imputation correctly predicted missing values and to test if the results obtained are valid. A common way to filter out false-positives is by running imputation over many iterations. Afterwards, SNPs are ranked by the number of iterations in which they appear and the most frequent SNPs are retrieved. Next, one can take a population level approach by looking at the distribution of sampled SNPs in each individual and the estimated allele counts that result from averaging each SNP occurrence over each iteration, using an r^2 type metric (Y. Li et al. 2009; Marchini & Howie 2010). One dilemma with these commonly used metrics are that rare genotypes may be removed after many iterations. In an effort to capture these true-positive SNPs, one can break down each genome into smaller subsets

before imputation. Selecting a subset of SNPs when making quality assessments allows granular control to keep rare SNPs (Hoffmann & Witte 2015). As frequentist association tests are known to inherently lose rare SNPs, overall study design to account for rare SNPs is crucial (Pei et al. 2010). Understanding population structure when collecting information is also paramount to fine-mapping these SNPs (Marchini & Howie 2010).

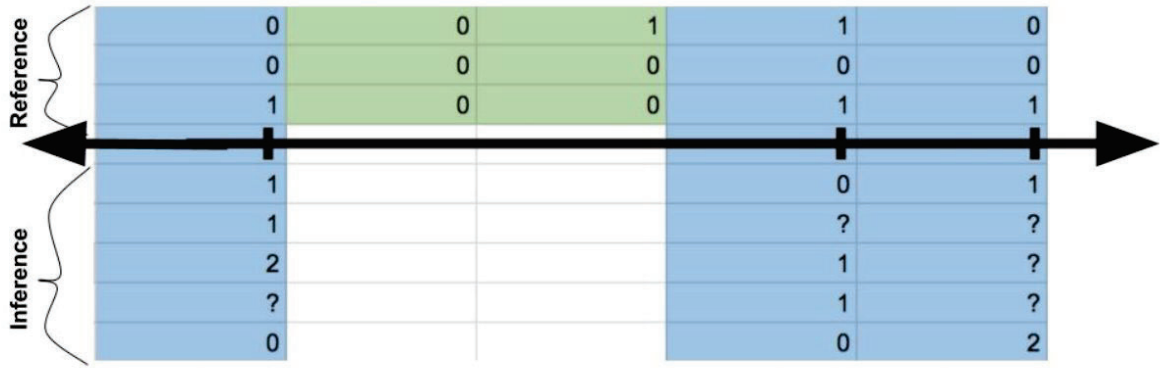


Figure 1.2 Imputation in GWAS General procedure for imputing genotype information. First, samples are phased at the genotyped SNPs. Next, SNPs are compared to the reference haplotypes (ex. HapMap 3 (The International HapMap 3 Consortium et al. 2010)). Sample haplotypes are matched to the best mixture of the reference haplotypes available. Finally, values for missing SNPs are predicted (McCarthy et al. 2016).

1.3.2 Imputation can be applied to epigenetic data

Imputation can also be applied to less traditional applications such as profiling of chromatin accessibility or chromatin marks. ChromImpute is a software designed to impute epigenetic marks (Ernst & Kellis 2015) with the aim of alleviating some of the issues that arise when obtaining epigenetic information. These issues include the financial and time cost, as well as not being able to fully map every histone mark in every tissue/cell type, which often leads to lack of power due to insufficient coverage (Rivera & Ren 2013). Additionally ChromImpute can be used to correct noise from general chromatin assays. The ChromImpute method is based on the fact that histone marks tend to be correlated such that the signal from different marks can be used to impute signal from histone marks that are missing or have errors (Ernst & Kellis 2010).

Given a large reference of histone marks compiled across a diverse set of tissues and cell types, ChromImpute provides a platform for imputation of various epigenomic signal tracks based on the data generated within an individual

experiment. Signal tracks represent the read coverage across a genome. These are broken up into 25 base pair (bp) bins genome-wide and each signal is averaged within the bins to provide an abstraction of the value. Because of the large number of assayed tissue types and histone marks, the ROADMAP and BLUEPRINT projects provide the most complete reference panels for ChromImpute (Ernst & Kellis 2015).

The software prioritizes a set of “main” histone marks (Tier 1 marks) which have the most observed data in the reference panel. These marks constitute the basis for imputation and consist of H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3, H3K27ac, H3K9ac and DNA accessibility (Ernst & Kellis 2015).

One major difference arises when evaluating broad and narrow histone modifications. Broad modifications generally span the entire gene range whereas narrow modifications seem to only span transcription factor binding sites (Starmer & Magnuson 2016). Additionally, when studying broad histone marks the obtained signal can be low even at large sequencing depths (Jung et al. 2014).

As the nature of imputation in ChromImpute is very different to that of genotype imputation software, the metrics for evaluation are also different. ChromImpute takes advantage of five key metrics to evaluate the overall efficacy of imputation: 1) a genome wide correlation between imputed (predicted) and observed data, 2) percentages of the top 1% 25 bp-bins of observed data that are also in the top 1% of the imputed bp-bins when ranked by signal intensity, 3) the percentage of the top 1% 25 bp-bins of imputed data that are also in the top 5% of the observed bp-bins when ranked by signal intensity, 4) the area under the curve (AUC) of a receiver operating characteristic curve when recovering 25 bp-bins in the top 1% of the observed data after ranking on imputed signal, and 5) the AUC when recovering 25 bp-bins in the top 1% of the imputed data after ranking on observed signal.

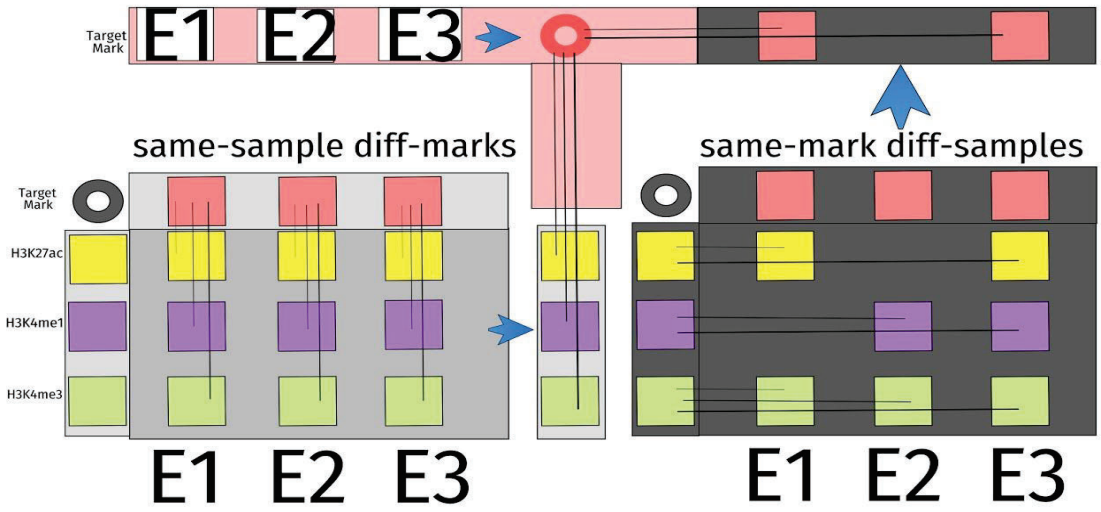


Figure 1.3 ChromImpute method This software begins by taking different samples (e.g. E1, E2, E3) and tabulating each of them based on the different marks within a given sample and then finding other samples with the same mark. Once that process has finished regression trees are built around the different marks assayed in a given sample. Additionally, the method utilizes the different samples for a given histone mark. By taking an ensemble based approach (combining the different regression trees) to calculate predictions, the algorithm does not have to worry about taking in any information from the target mark. After each regression tree is combined the imputation for a given histone mark for a given sample can be calculated using ChromImpute (Ernst & Kellis 2015).

1.4 Thesis outline and goals

The main objective of this thesis is to perform and evaluate the imputation of epigenetic marks in a rare cell type (Tregs) at large scale. Given the importance of Tregs in autoimmune diseases, it is becoming markedly clear that the different genomic data generated holds valuable insight into the biology of the cells. Using statistical techniques like imputation to improve data recovery and quality holds a crucial importance in further studies. Using ChromImpute, I imputed several Tier 1 histone marks in Treg cells and analysed the results to determine the advantages and limitations of routinely implementing imputation in ChIP-seq experiments.

Aims

- To evaluate if imputation can improve ChIP-seq reference catalogues while preserving inherent data structure
- To assess whether imputation can be used in ChIP-seq analysis of genetic variability

2. Methods

2.1 Samples

The following protocol and design were carried out by Dafni Glinos, a PhD student in the lab and Dr. Natalia Kunowska, a Senior Research Assistant in the lab.

1. Setup: The ChIPmentation-seq (ChM-seq) protocol (Schmidl et al. 2015) was performed on 100,000 sonicated Treg cells. After sorting and isolating Treg cells, those same cells were resuspended in full medium (IMDM, 10%FCS) at 1-2 million cells per ml.
2. ChIP-seq protocol:
    a. ChIP-seq libraries were generated using Illumina TruSeq index tags, while ChM-seq libraries were generated using the Nextera dual index tags.
    b. They cross linked cells by resuspending in a 1% formaldehyde solution.
    c. Following a 5 minute incubation at 37℃ the cells were quenched with glycine for 5 minutes to achieve a 125 mM concentration.
    d. They lysed and sonicated cells using the iDeal ChIP-seq kit for histones (Diagenode) as instructed by the manufacturer.
    e. They performed immunoprecipitation (IP) using the same kit. For this step, they used ChIP-seq grade antibodies against the histone marks H3K27ac (Diagenode), H3K27me3 (Abcam), H3K4me1 (Active Motif) and H3K4me3 (Active Motif).
    f. They prepared sequencing libraries as instructed by the iDeal ChIP-seq protocol. When doing ChIPmentation-seq, the two previous steps were combined by adding the Nextera Tn5 transposase after the IP step.

3. Sequencing: Samples were loaded for sequencing on a HiSeq 2500 instrument and V4 chemistry using standard 75 bp paired-end reads. In total there were 14 ChIPmentation-seq libraries that averaged 62 million reads per sample.
4. Data pre-processing: After data acquisition, the Sequencing Facility at the Wellcome Sanger Institute demultiplexed the sequencing data and mapped it to the human genome. Sequencing files were released in CRAM format.

2.2 ChIP-seq data processing

I re-mapped the reads from build 37 (GRch37) to the human reference genome build 38 (GRCh38) using bwa mem algorithm (Li 2013), and converted the aligned reads into BED (Kent et al. 2002) format. Only uniquely mapped reads were kept and duplicates were removed using samtools version 1.3.1 (H. Li et al. 2009).

The mapped reads were built into a signal track in bedgraph format using MACS2 (Zhang et al. 2008). The method uses input data i.e. control data, which is generated by treating samples the same way in the ChIP-seq protocol but without the addition of the antibody against the target epitope. The method then computes if the local average read coverage is statistically different compared to the control background. This data is referred to as observed data. This step is meant to render the output signal tracks to correct for any biases in the different samples.

2.3 Imputation reference panel

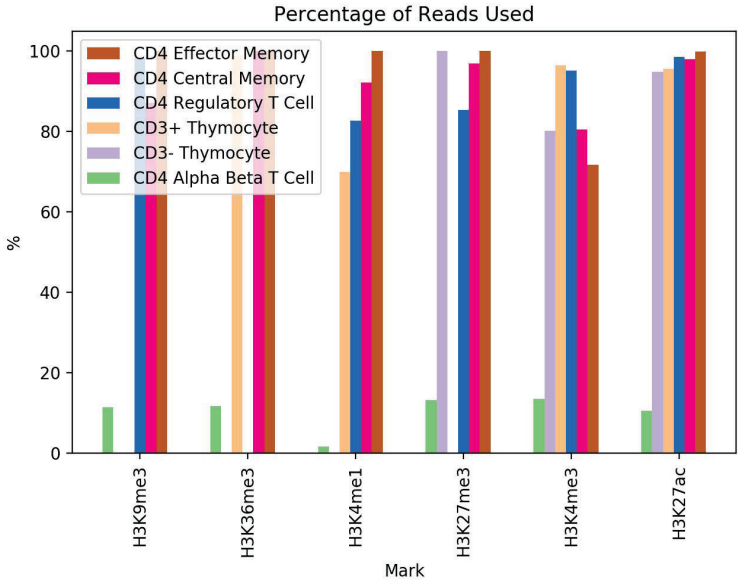
ChromImpute takes the information from the reads as input and converts it into smaller fragments, in the *Convert* step. Specifically, during the *Convert* step the software takes the signal track and averages the read coverage over 25 bp bins across the genome.

The construction of the reference panel was performed in two steps. First I built the reference panel with the Roadmap (Bernstein et al. 2010) and ENCODE data

(ENCODE Project Consortium 2012; ENCODE Project Consortium 2004) as specified in the ChromImpute paper (Ernst & Kellis 2015). For that I used 129 cell types and 8 histone marks: DNase, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9ac, H3K9me3. Second, I added the BLUEPRINT (Adams et al. 2012) data to the reference. For that I selected cell types within the general T cell population: $CD4^+$ Effector Memory, $CD4^+$ Central Memory, $CD4^+$ Regulatory T cell, $CD3^+$ Thymocyte, $CD3^-$ Thymocyte, $CD4^+$ Alpha Beta T cell. I used any given mark that was available, and the marks that I used were: H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9ac, H3K9me3.

The number of reads for each of the cell types varied greatly; most of the reads came from $CD4^+$ Alpha Beta T cells as these represented the vast majority of the total samples. I aimed to have between 100 to 160 million reads per mark per cell type as that was the general baseline required to prevent any bias towards a specific cell type/mark combination. If there were more than 160 million reads available, I would randomly subsample the different read files to reach the total of 160 million reads. For example, the $CD4^+$ Alpha Beta T cells had 31 samples with over one billion reads. Thus the percentage of reads sampled was far less than for other cell types **(Figure 2.1.A)**. Overall, I constructed the reads based on what was available; many cell types had a lower number of reads (Tregs, Effector Memory). If there were less than 160 million reads in the entire cell type, I used them all in the reference **(Figure 2.1.B)**.

A) Percentage of reads used



B) Number of reads sampled

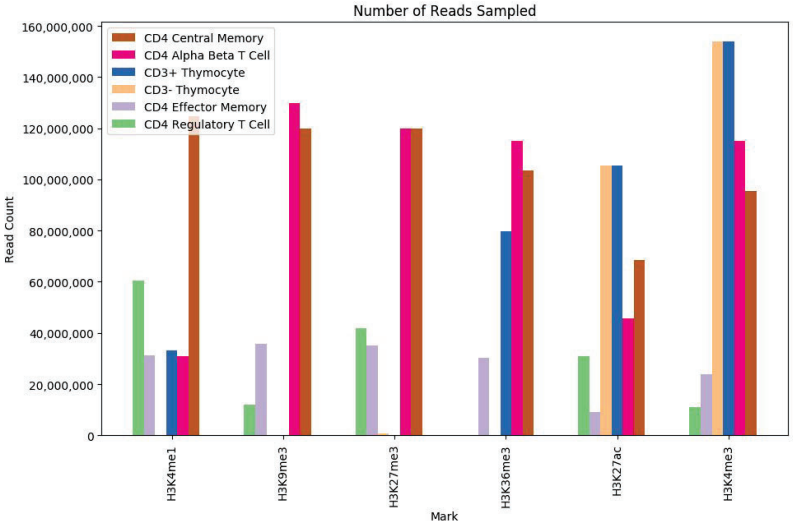


Figure 2.1 Reads assembly broken down by cell type and histone mark The Blueprint reference panel reads were assembled using downloaded FASTQ (Cock et al. 2010) files from: https://www.ebi.ac.uk/ega/user/login. The files were then converted into

BAM (H. Li et al. 2009) format and combined and added to the existing consortium. **A)** describes the percentage of reads used in the reference and **B)** illustrates the total number of reads sampled in each histone mark.

2.4 Imputation

There are four steps in the imputation pipeline. The first step is to compute the global distance between each of the observed data signal tracks and the reference signal tracks for the same histone mark, using a genome wide signal correlation between the imputed and observed tracks. For this, I used the *ComputeGlobalDist* method in ChromImpute. The obtained output was a list of samples ranked by the correlation with the observed sample signal track in the reference panel for the given histone mark. The second step, *GenerateTrainData*, generates training data by using the signal tracks from the same sample with different marks. This step is meant to capture any specific sample information and requires that at least 2 histone marks are assayed for any given sample. The training data is then combined with the most similar samples, up to ten samples, based on the same histone mark. In the third step, the model is trained using regression trees, this corresponds to the *Train* command. Each of the different signal tracks are transformed into a regression tree that is used to compute imputed signal at a given position. Finally in the fourth step, the different trees are then applied, *Apply*, onto the observed signal track and using an ensemble approach, the information from each tree is combined and averaged. The final result is a signal track with imputed signal. There is no correction for read depth or any biases, and samples that may be deeply sequenced may impute in a different way than samples with low sequencing depth.

2.5 Evaluating reference panel

In order to understand the effectiveness of the reference panel, I performed a signal track recovery analysis to evaluate which marks performed best during imputation. The marks with the best recovery were H3K27ac, H3K4me3 and H3K4me1 as there is the most reference data for these marks. For that I use the *Eval* tool, provided by

the ChomImpute. The output of this analysis are several metrics for each sample and histone mark combination, including: the fraction of the observed top *x* percent signal track in the imputed top *x* percent signal track and vice versa; the genome wide correlation between the observed and imputed signal tracks; the area under the ROC for predicting the top *x* imputed signal with the observed signal; and the area under the ROC for predicting the top *x* observed signal with the imputed signal. Here *x* represents a value that a user can specify for ChromImpute.

2.6 Downstream computation of observed and imputed peaks

In order to call peaks, and generate the initial signal track, I used the MACS2 (Zhang et al. 2008) software. To generate signal tracks I used the commands: *macs2 callpeak -t <index_file> -f BEDPE -n <temp_peaks_file> -g hs --nomodel -B --SPMR* along with *macs2 bdgcmp -t <bdg_reads> -c <bdg_input> -o <bedgraph_output> -m ppois -S <s_value>*. In order to call narrow peaks I used this command: *macs2 bdgpeakcall -i <bedgraph_input> -c <cutoff> -l <min_length> -g <max_gap> --outdir <output_dir> -o <output_peak_file>*. To call broad peaks use the same command but swap *bdgpeakcall* with *bdgbroadcall*.

I used BEDTools (Quinlan & Hall 2010) to analyze the generated peaks. To compare which peaks were shared as well as unique to the imputed and observed datasets I used the *intersect,* command. An example command to find 20 percent overlap in two sets of peaks is: *bedtools intersect -a Observed.bed -b Imputed.bed -f 0.2 -r -wo > 20percentInBoth.bed*. To convert BAM files to BED format I used the following command: *bedtools bamtobed -bedpe -i <bam_file> > <bed_file>*.

I used ChIPSeeker (Yu et al. 2015) to provide gene annotations and follow up with functional analysis on the different peak files. I also used the UCSC known genes version 3.4.0 to annotate genes (Hsu et al. 2006). Exentensive documentation can be found here: *https://bioconductor.org/packages/release/bioc/vignettes/ChIPseeker/inst/doc/ChIPseeker.html*.

I used several Python libraries for analysis: Pybedtools (Dale et al. 2011) as an API for BEDTools, Pandas (Mc Kinney n.d.) for data table manipulation, matplotlib (Hunter 2007) and seaborn (Waskom et al. 2014) for plotting and graphs, PyPlink to handle PLINK (Purcell et al. 2007) files with respect to genotype data, Pathos (McKerns et al. 2012) for multiprocessing and code optimization, NumPy (Oliphant 2006) and SciPy (Millman & Aivazis 2011; Oliphant 2007) for statistics and general mathematics. I used iPython (Perez & Granger 2007) and Jupyter Notebooks (Kluyver et al. 2016) for quick code prototyping as well. I used Pandas to handle all of my data, which allowed for ease to use relational database concepts with tables loaded in RAM (Random Access Memory). I used SKlearn (Pedregosa et al. 2011) to perform a Linear Regression as well as normalization within the generated Pandas dataframes.

I used the UCSC Genome Browser (Kent et al. 2002) to manually verify peak overlap and compare and contrast observed and imputed peak information.

The code for the ChromImpute pipeline along with every Python script I developed for analysis is available at:
https://gitlab.internal.sanger.ac.uk/TrynkaLab/ChromImpute.

3. Results

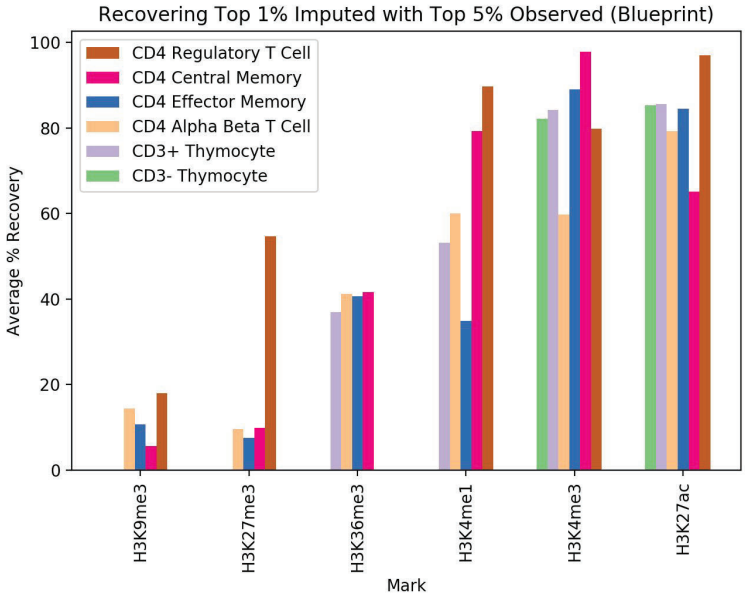
3.1 Overview

In this study I focused on understanding how imputation can augment different experimental ChIP-Seq assays. In this chapter, I will explore the results of the global structure of chromatin when ChromImpute is applied to Treg data, over various histone marks. I will then examine the impact of ChromImpute on the technical variability of the ChIP-Seq assay. Additionally, I examine the best MACS2 peak parameters to use when working with imputation data. Finally, I describe the effects of imputation on genotypic variance when analyzing a selection of peaks with eQTL effects.

3.2 Imputation preserves ChIP-seq data structure globally

To determine whether ChromImpute preserves the global structure of chromatin data in addition to reducing noise, I examined ChIP-seq data from 3 samples of regulatory T cells generated by the Trynka lab. The histone mark assayed was H3K27ac. I used the H3K27ac mark as the basis for our evaluation as this mark was the most complete and had the best recovery for Tregs in the BLUEPRINT and the Trynka Lab samples **(Figure 3.1)**. Peaks called with a p-value lower than 10^{-5} were analyzed using ChIPseeker. I use the term “observed” when referring to the data before any imputation had been performed on it.

A) Recovery of base pair bins in Blueprint samples



B) Recovery of base pair bins in Trynka Lab samples

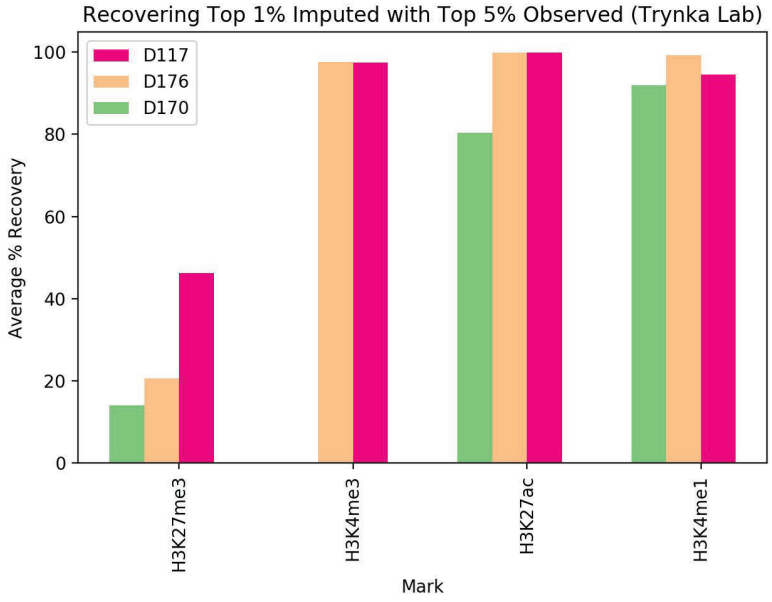


Figure 3.1 Imputation accuracy This metric evaluates the top one percent of signal bins in the imputed data that is also in the observed data. The Blueprint samples, **A)** showed the highest recovery in H3K27ac and H3K4me3 marks reads used in the reference. Similarly, the Trynka lab samples **B)** showed the most recovery in H3K27ac, H3K4me3 and H3K4me1

marks. Recovery is defined as how much of the top observed signal track is preserved in the imputed signal track.

I found that more peaks overlap with transcription start sites (TSS) in imputed compared to observed data **(Figure 3.2)**. Observed samples showed significant variability in read count frequency. For example, D170 had the highest read count frequency and the lowest sequencing depth. However, upon imputation all samples showed a similar distribution around the TSS.

Next, I investigated the location of imputed and observed peaks. I performed this analysis on sample D176, which had the highest sequencing depth. I observed an increase in peaks located in proximal promoters (within 1kb of the TSS) in imputed compared to observed data **(Figure 3.3)**.

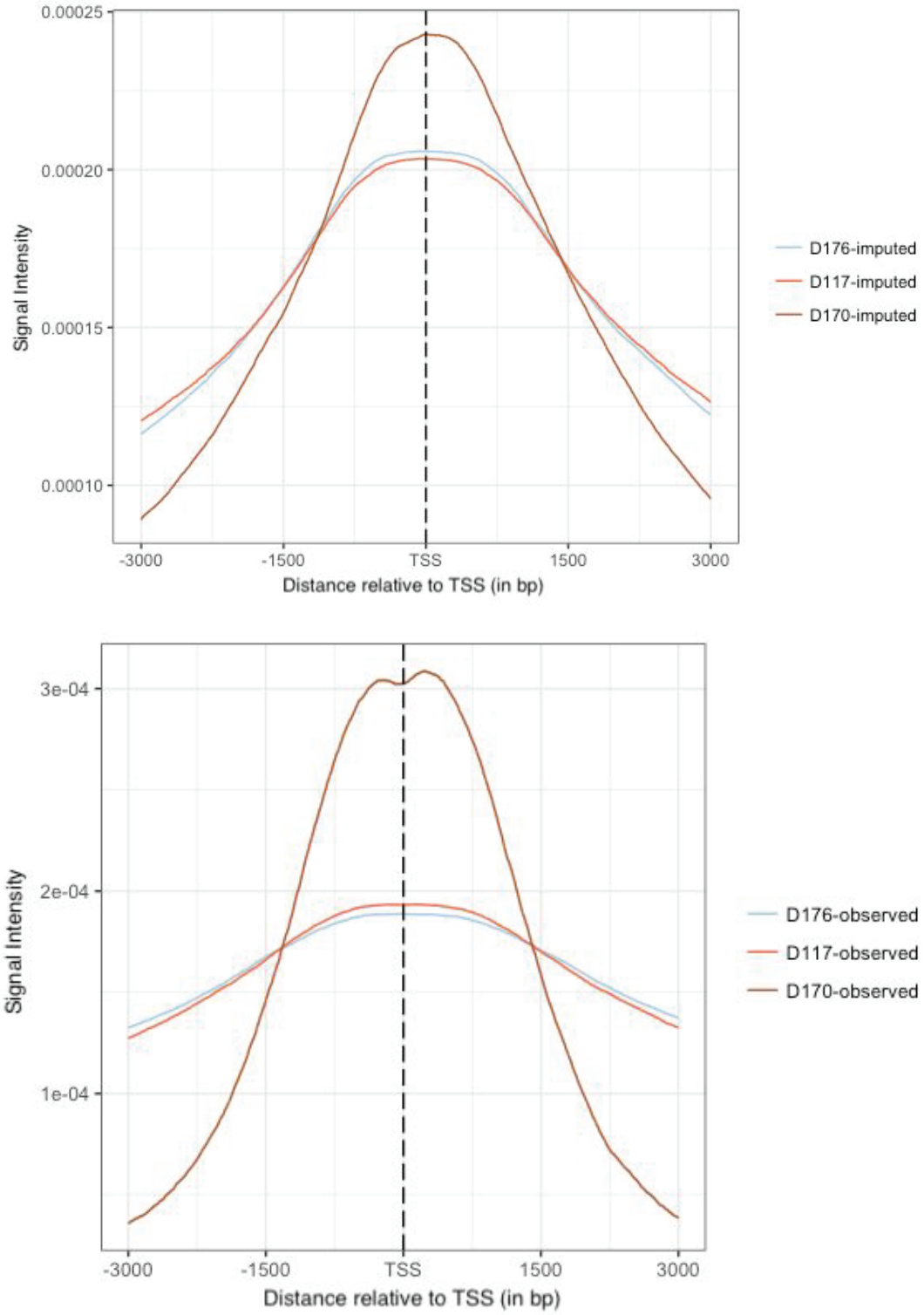


Figure 3.2 Peak overlap with TSS I overlapped peaks to the nearest TSS position within +/- 3kb window. The imputed peaks overlapped closer to the TSS. Results correspond to each respective sample (color code) for the H3K27ac mark. The top panel contains the imputed peaks, while the bottom contains the observed peaks.

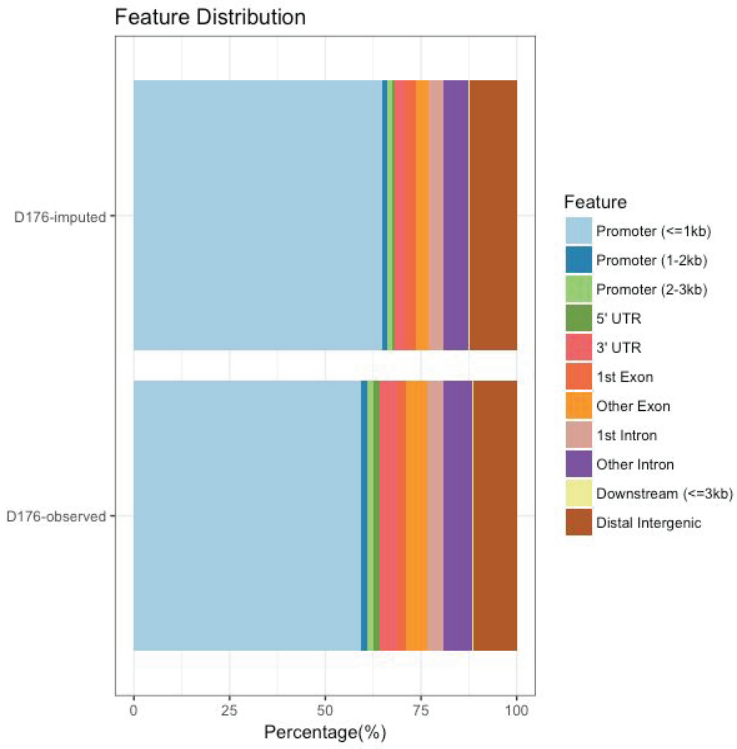
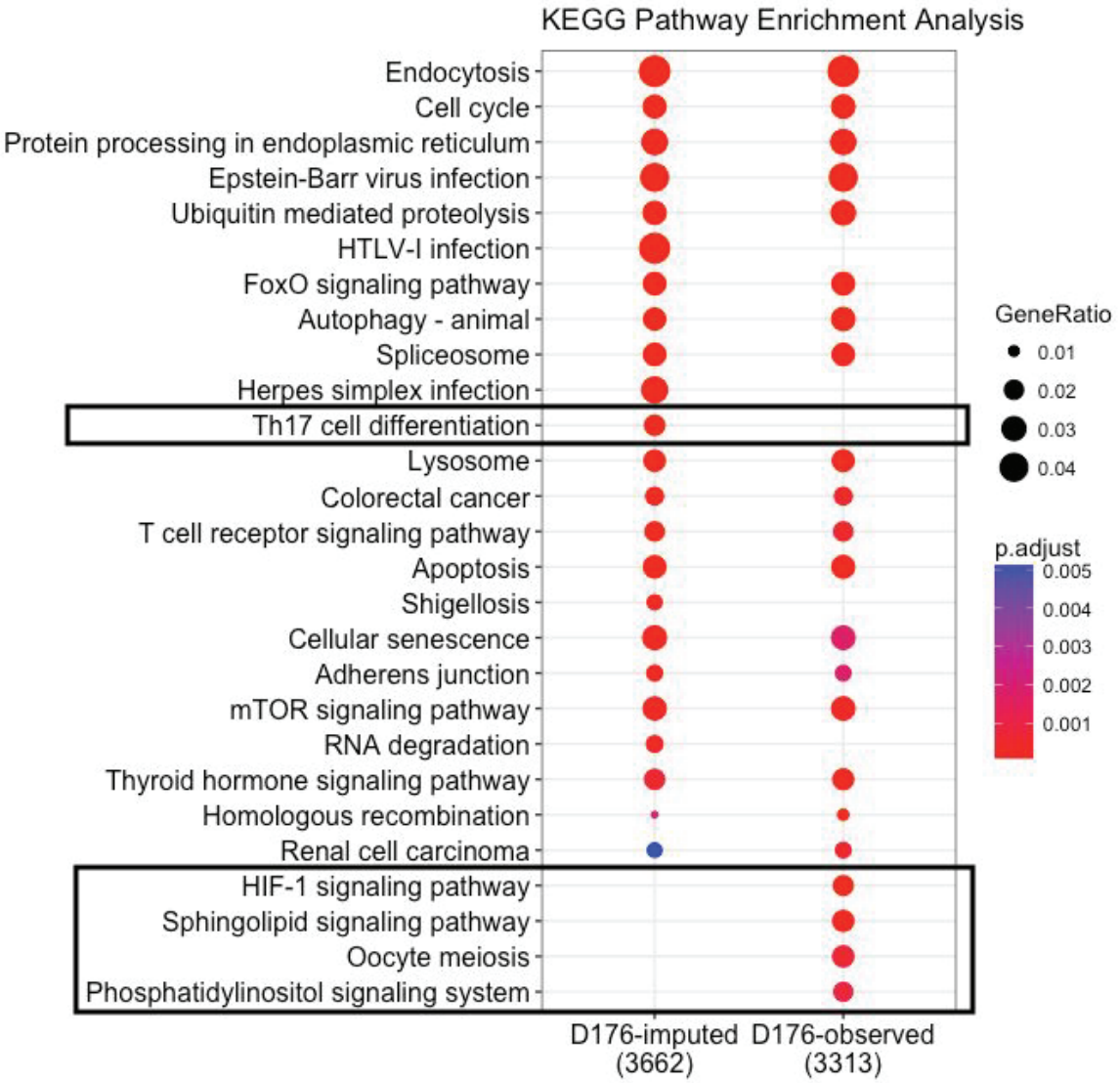


Figure 3.3 Peak annotation relative to gene features. Peaks were annotated by distance to a TSS using UCSC HG 38 known genes (Hsu et al. 2006). When there were multiple peak annotations, the annotation within closest distance to the TSS was chosen. This distribution was able to capture more promotor peaks in the imputed data, but in doing so neglected to capture as much enhancer data. This may clean up spurious peaks, but in the process re-align them with promoter regions.

A) Pathway enrichment for H3K27ac comparing imputed vs observed peaks



B) A set of randomly associated peaks for H3K27ac

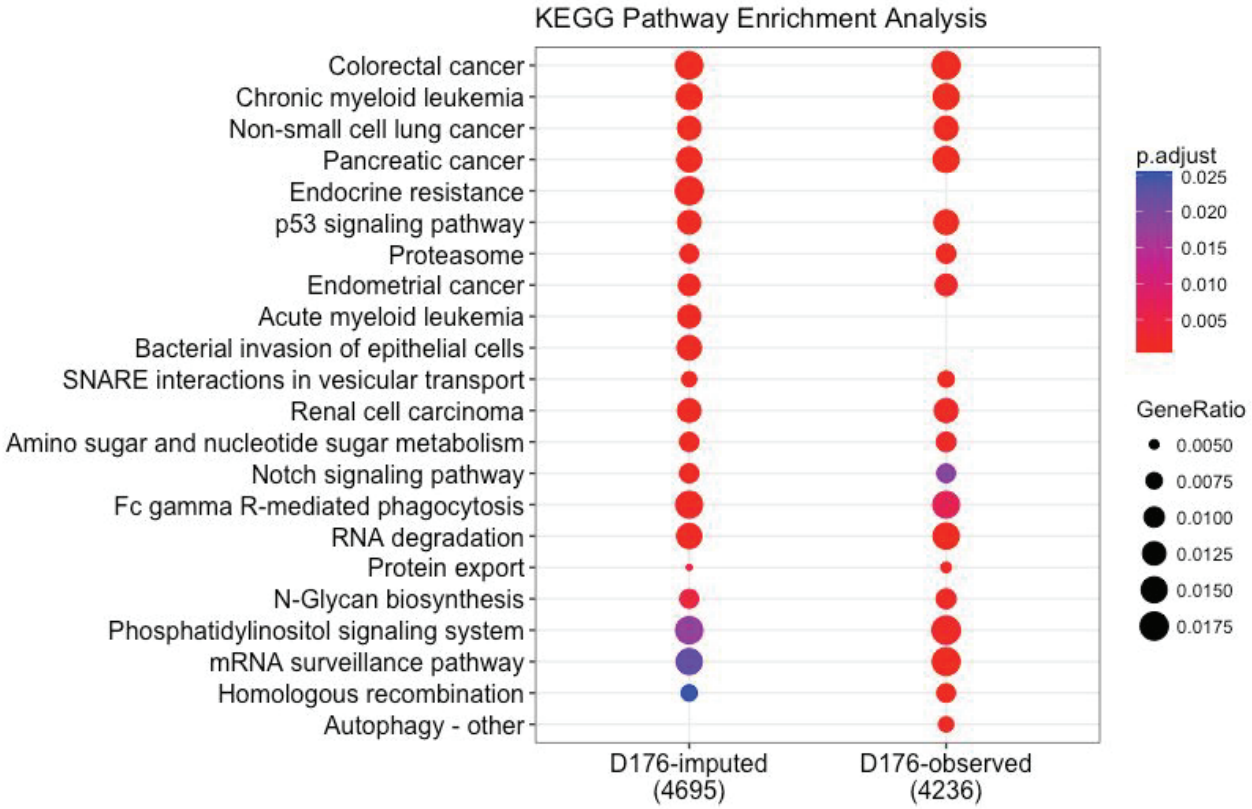


Figure 3.4 KEGG Pathway Enrichment analysis Genes within 3kb to ChIP-seq peaks were annotated via pathway enrichment analysis using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Circle areas represent the gene ratio, which is proportional to the number of genes that overlap a specific pathway. Colors represent the adjusted p-value of enrichment for the respective gene set. The KEGG pathway in **A)** highlights the Treg specific pathways found in contrast to **B)** a set of random peaks that were picked. This was analyzed for the H3K27ac mark. The imputed pathways seemed to be more focused to the biology of the T cell.

In order to understand if imputed peaks were relevant to the biology of Treg cells, I performed pathway enrichment analysis using KEGG **(Figure 3.4)**. I observed noticeable additions in the imputed Treg peaks, including Th17 differentiation. Pathways that were dropped included Oocyte meiosis, Sphingolipid and HIF-1 signaling pathway, and Phosphatidylinositol signaling system. However, the majority of pathways are preserved between the two datasets. I concluded that imputation

eliminated noisy peaks, while preserving the inherent characteristics of active chromatin in Tregs. The noisy peaks annotate to pathways that are not necessarily specific to T Cells and those false peaks are removed based on the distribution of the peaks on the specific gene annotations.

3.3 Imputation reduces technical variability in ChIP-seq data

I was interested in understanding the effect of ChromImpute on the technical variability of ChIP-seq assays. Thus, I imputed ChIPmentation data from 11 Treg samples and two histone marks: H3K4me1 and H3K27ac. My assumption was that, since samples were from the same cell type and chromatin mark, they should be similar in signal track structure and contain a similar set of peaks. I observed that the observed and imputed data had on average 10,100 and 13,600 number of peaks, respectively. However, imputation markedly reduced the variability in number of peaks compared to observed data **(Figure 3.5)**. I performed Bartlett's test for equal variances which returned a statistically significant value to indicate that the variances between the number of peaks were indeed different.

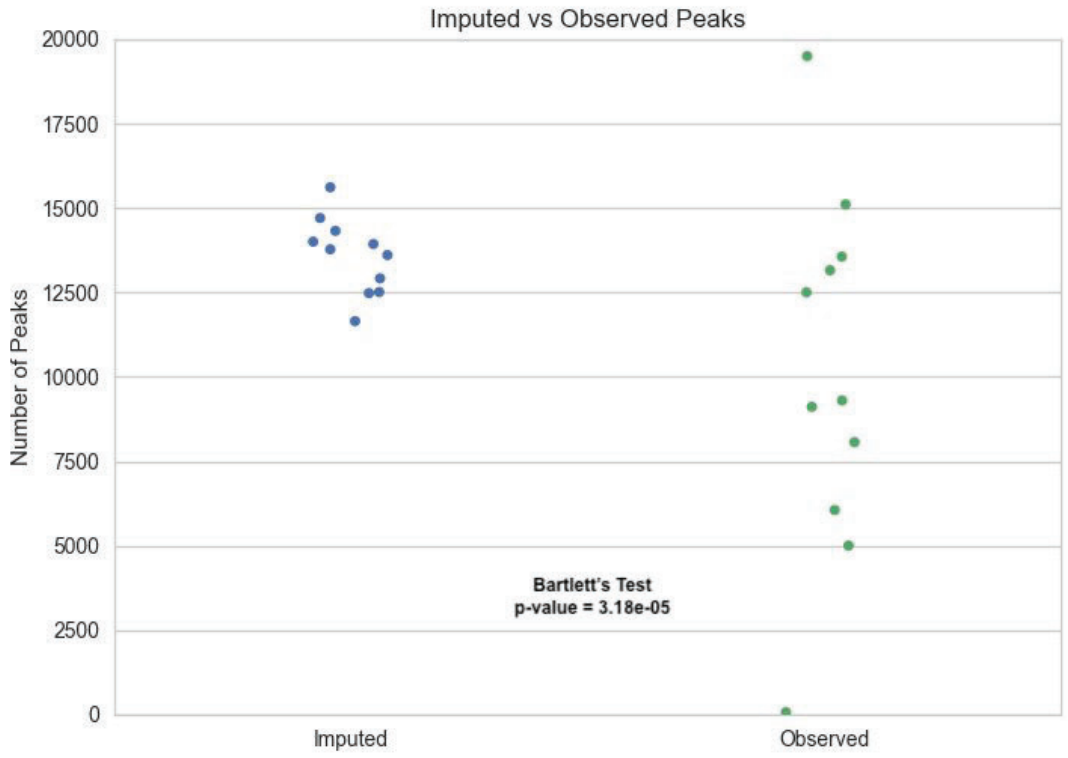
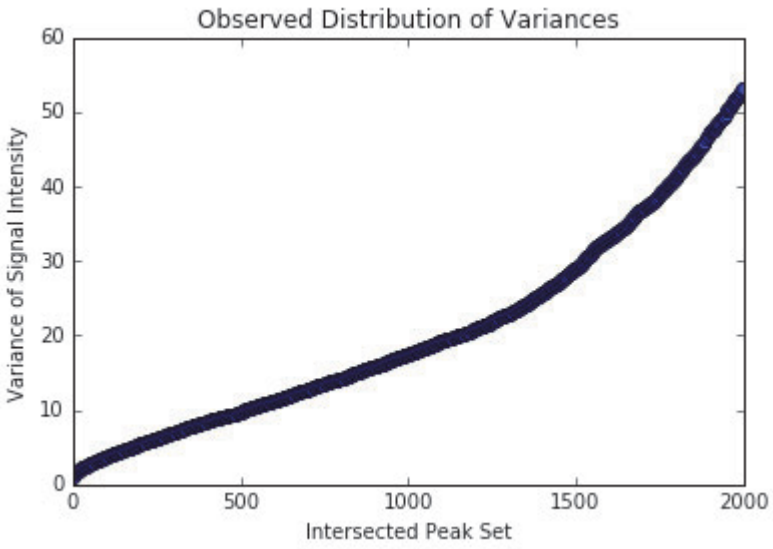
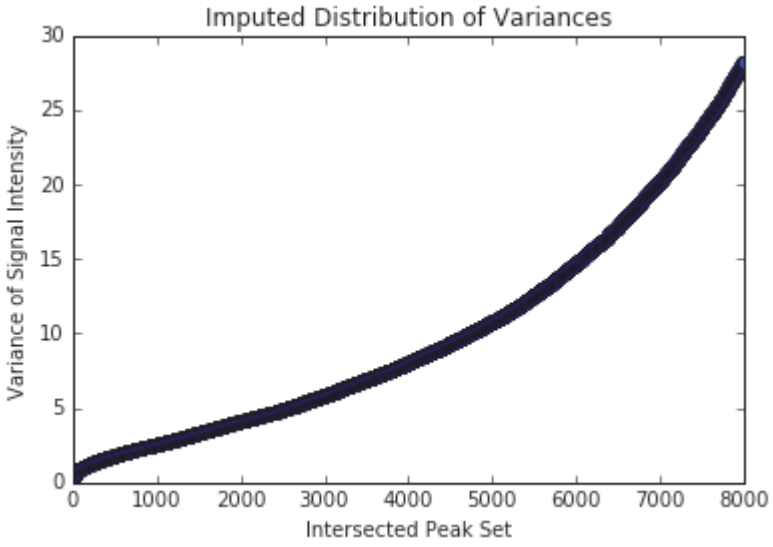


Figure 3.5 Comparison of variance in imputed and observed data. The total number of peaks per sample called across 11 ChIPmentation samples for the observed and imputed data. Bartlett's test for equal variance revealed a p-value of 3.18×10^{-5}.

I then asked if the imputed and observed data contained the same set of peaks, as well as how much signal intensity varied in observed and imputed peaks. I intersected the 11 imputed samples and the 11 observed samples using an overlap threshold of 20% between two features. This resulted in a common set of peaks, detected in both observed and imputed data. Since each peak had an associated signal intensity, I then calculated the variance in peak signal intensity across biological replicates in both data sets. I sorted each of the peaks by variance to visualize the differences between the two sets of peaks. To control for outliers I analyzed, 80% of the shared peaks **(Figure 3.6)**. I observed that signal intensity varied 50% less in imputed than it did in observed peaks.

A)


Observed Distribution of Variances
Variance of Signal Intensity
Intersected Peak Set
0
10
20
30
40
50
60
0
500
1000
1500
2000


Imputed Distribution of Variances
Variance of Signal Intensity
Intersected Peak Set
0
5
10
15
20
25
30
0
1000
2000
3000
4000
5000
6000
7000
8000

B)

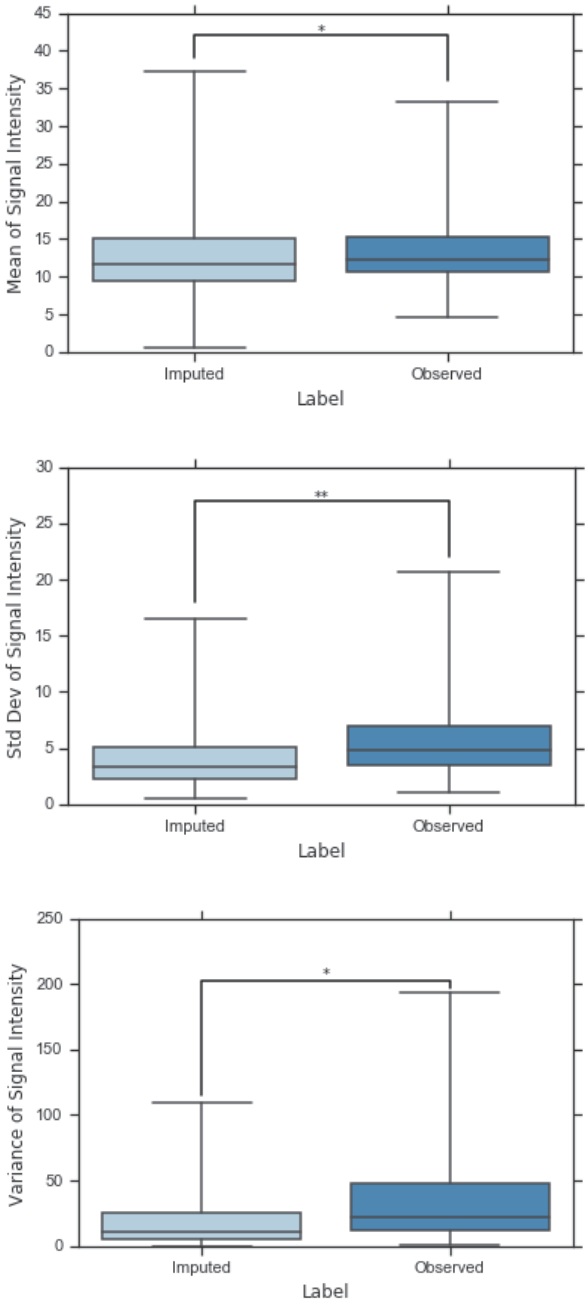
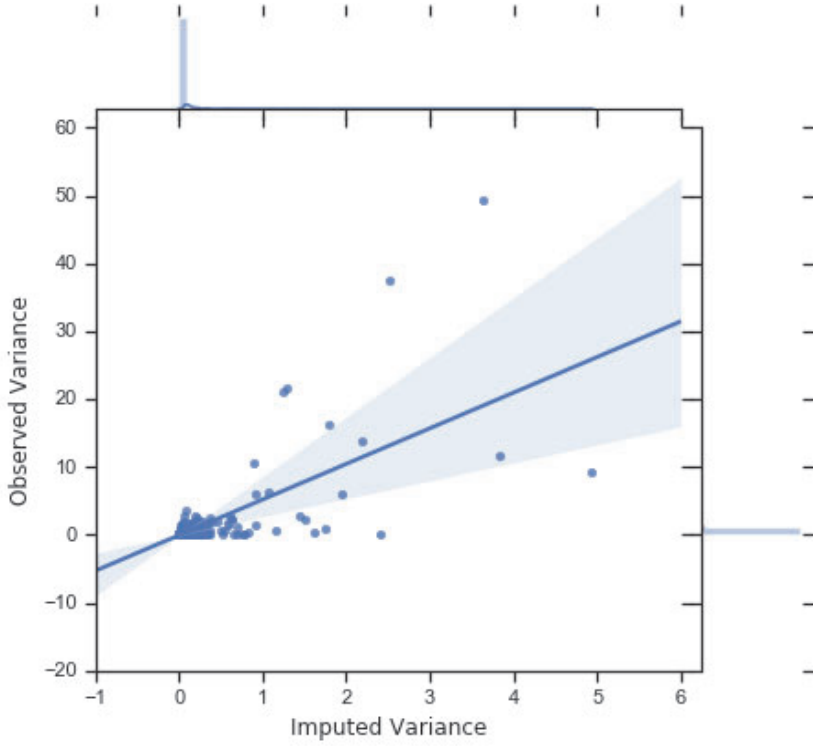
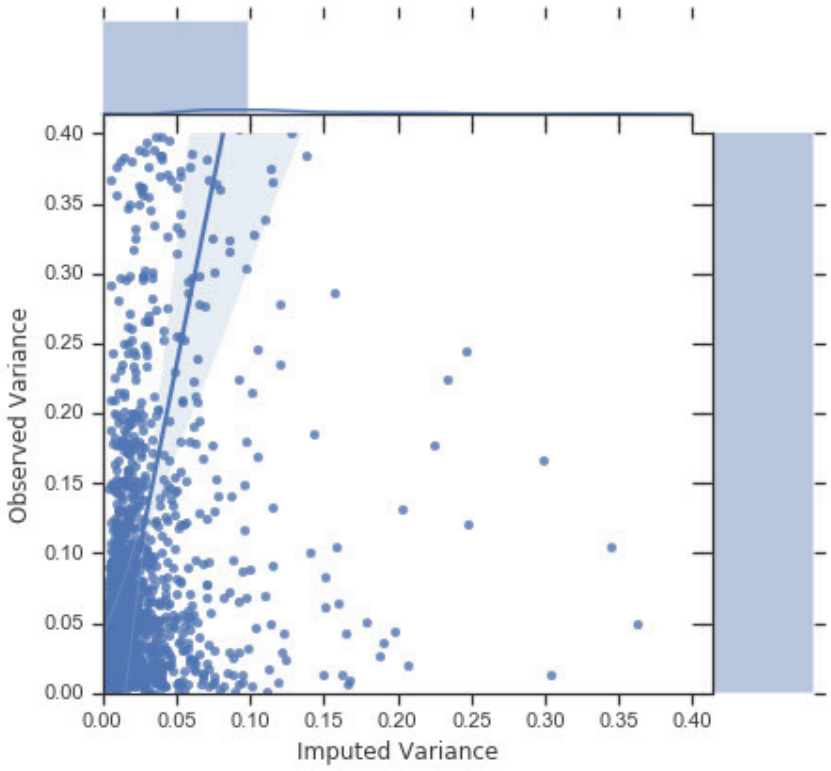


Figure 3.6 Distribution of variance for imputed and observed peaks. I identified peaks that were shared between observed and imputed data by intersecting peaks from 11 samples (minimum overlap of 20 percent). **A)** The distribution of peak variance for signal intensity for 80 percent of the peaks. **B)** I analyzed the variance of signal intensity to provide mean, standard deviation and variance. * indicates a p-value < 0.05 and ** indicates a p-value < 0.005 using a two sample t-test to calculate the p-values.

A)


Observed Variance
Imputed Variance
60
50
40
30
20
10
0
−10
−20
−1
0
1
2
3
4
5
6

B)


Observed Variance
Imputed Variance
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40

C)

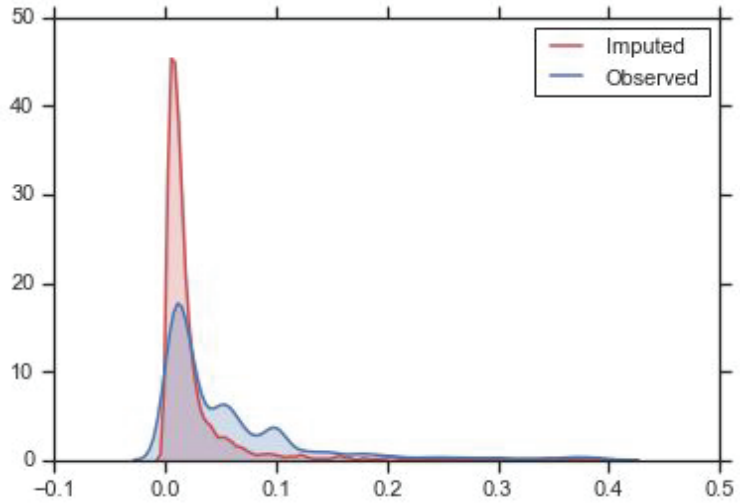
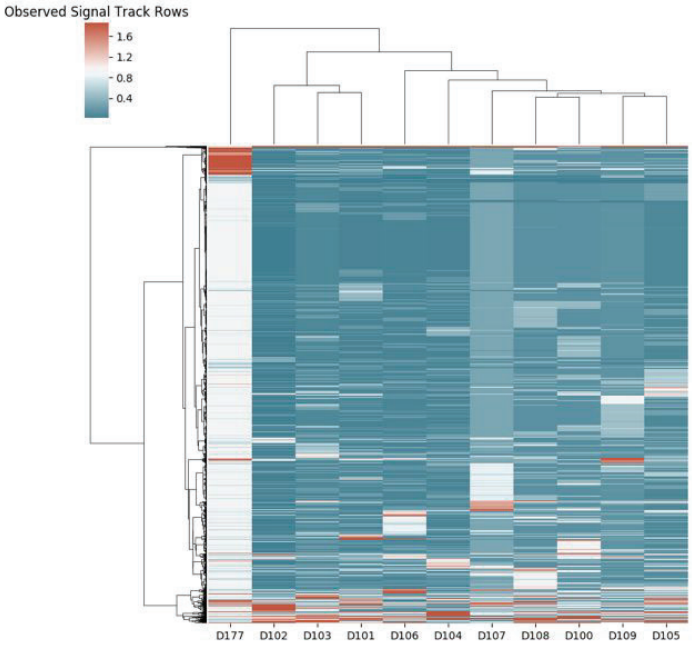


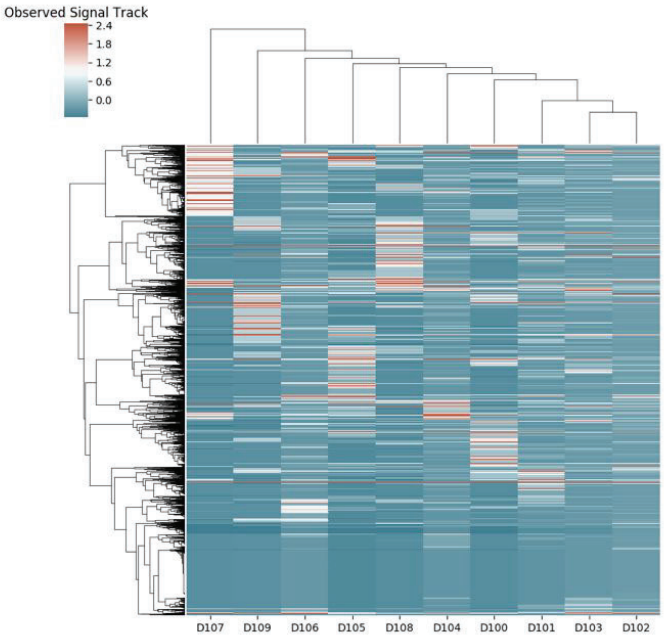
Figure 3.7 Variance between imputed and observed signal track I decomposed observed and imputed signal intensity tracks into 100,000 base pair bins and calculated the variance in observed and imputed signal intensity across the 11 biological replicates. Variance was calculated using the same bin at the same genome position for the observed and imputed samples (y- and x-axis respectively) **A)**. The zoomed in plot **B)** captures the variance difference between imputed and observed. Marginal density plots **C)** display regions of high density within the observed and imputed data for the zoom in.

I wanted to understand what was driving the differences between observed and imputed signal tracks. In order to answer this question, I decomposed the observed and imputed signal track for each sample into 100,000 base pair bins. The imputed signal tracks varied far less than the observed tracks **(Figure 3.7)**.

A) Observed



B) Observed without D177



C) Imputed

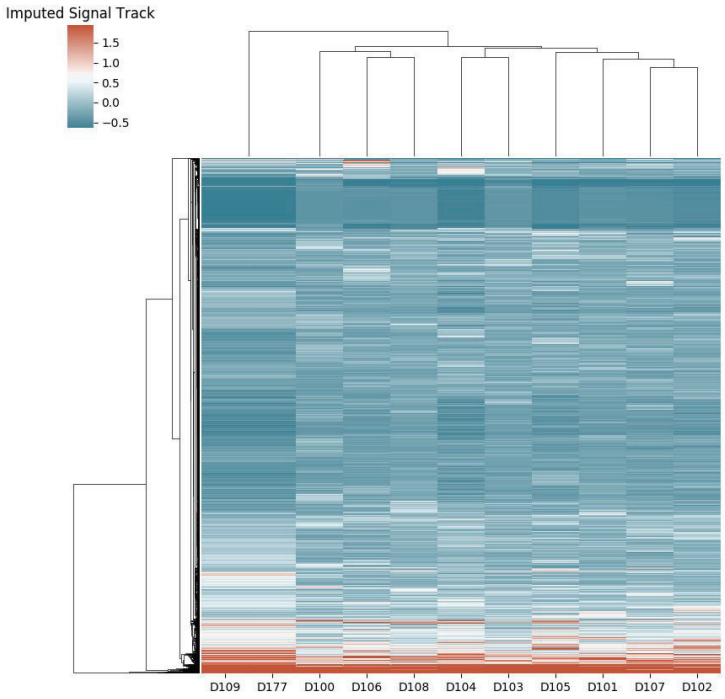


Figure 3.8 Heatmaps of Unweighted Pair Group Method with Arithmetic Mean clustered imputed and observed data. To visualize the similarity between the observed and the imputed data I performed Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clustering on signal intensity in 100,000-bp bins throughout the whole genome. **A)** All observed samples, including the low quality sample (D177, most left outlier sample); **B)** Observed samples excluding the low quality D177 sample; **C)** Imputed samples including the low quality D177 sample, which after imputation is no longer an outlier and clusters closely with the D109 sample.

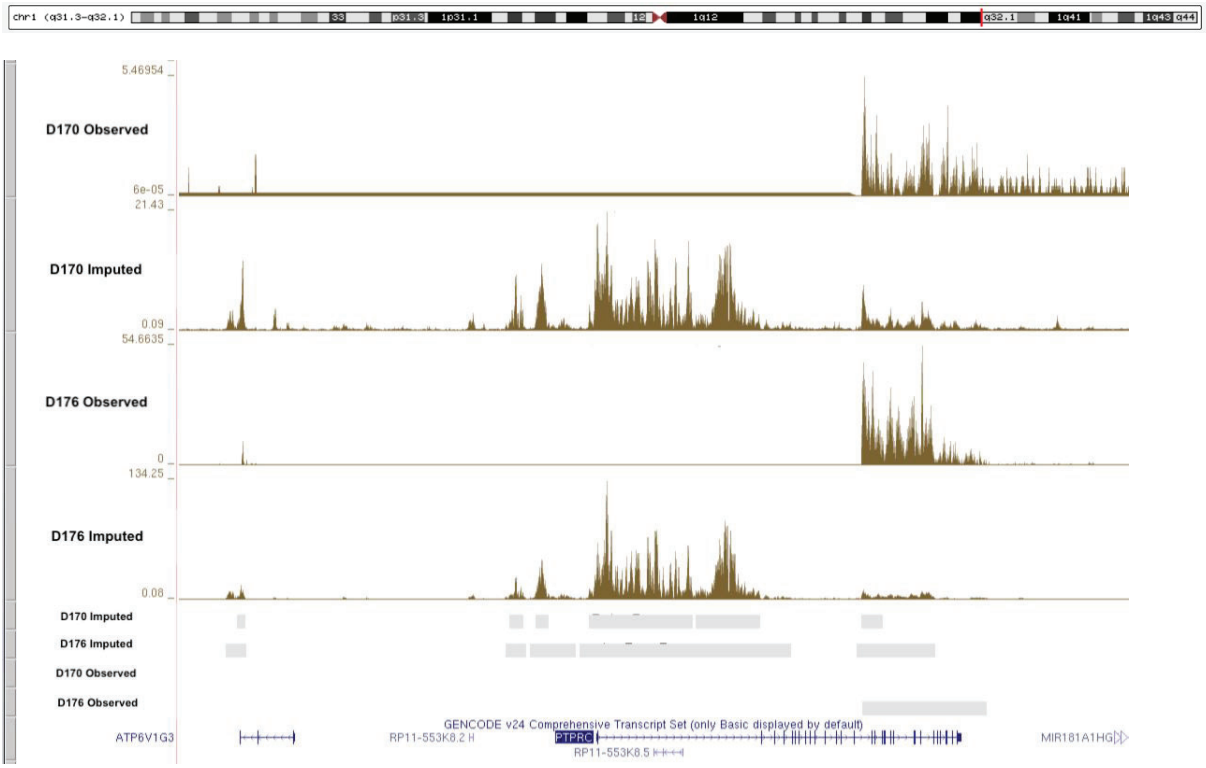
Finally, I asked if imputation preserved inter-individual variability. I organized observed and imputed signal tracks into heatmaps which included the 11 Treg samples previously analyzed **(Figure 3.8)**. The tracks were ordered by hierarchical clustering of Euclidean distances. I found that signal intensity varied more in the observed samples than it did in the imputed samples **(Figure 3.8.A)**. This inter-individual variation in observed samples was evident even after removing the

outliers (D177) **(Figure 3.8.B)**. Moreover, the dendrogram of samples obtained using the observed signal intensities showed a completely different order to the dendrogram of imputed samples **(Figure 3.8.C)**. Thus, I concluded that when ChromImpute is applied to a set of signal tracks from different individuals, the imputed tracks are much more alike and the inter-individual variation is lost.

3.4 Imputation corrects for experimental biases and missing data

Often times experimental errors can hinder different experimental assays, and often generate false results. I asked whether ChromImpute could help correct for these false results and prevent any type of introduced errors or missing data. When visualising the data, I noticed that two Treg ChIP-seq samples had lost any read pileup spanning the PTPRC gene **(Figure 3.9.A)**. Furthermore, other ChIPmentation samples processed did show signal over this gene **(Figure 3.9.B)**. When the data was imputed, ChromImpute was able to recover this signal **(Figure 3.9.A)**. The file was corrupted and did not include the expected signal. Thus, ChromImpute is able to fill in missing information and ultimately build strong reference signal tracks for any further analysis downstream.

A) Recovery of signal via imputation



B) Expected signal without any errors

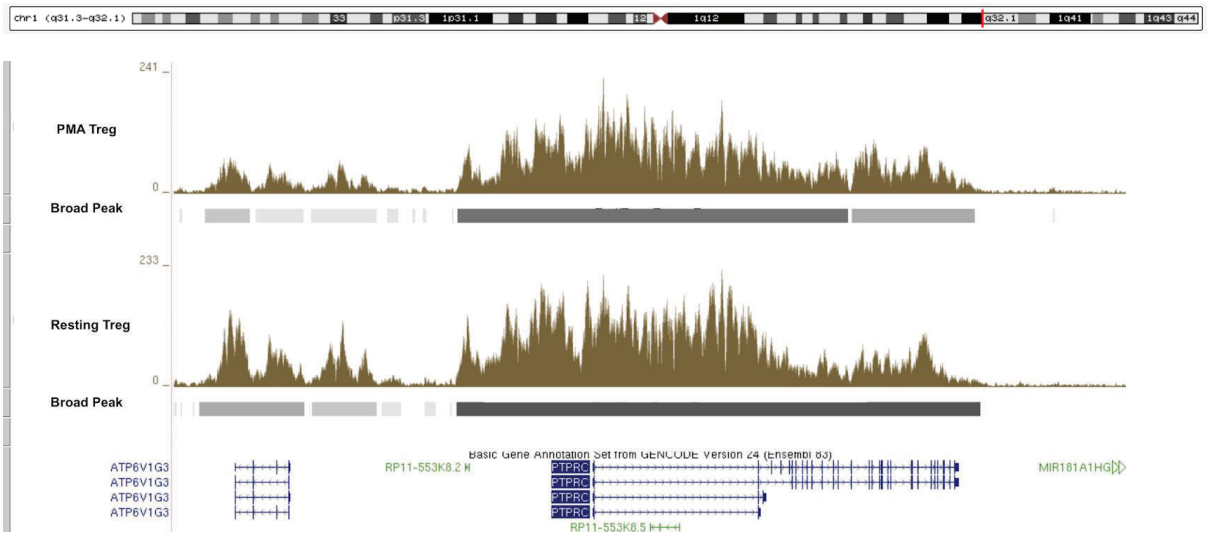


Figure 3.9 Recovery of signal through imputation. Samples D170 and D176 are two individuals for which I observed a loss of intensity signal over a Treg critical CD45 gene, PTPRC for histone mark H3K27ac. The location of this gene is at chromosome 1:198,639,040-198,757,283. **A)** Upon imputation the signal was recovered in both the

replicates. Moreover, when other Treg samples are processed, the signal is very evident **B)** indicating the error likely stemmed from a mishap in the assay.

3.5 Imputation data should be assessed with a broad peak caller

Peak calling provides a critical basis for evaluating regions of importance in various sequencing protocols. I wanted to evaluate how well the MACS2 peak caller performed on observed and imputed ChIPmentation data. In particular, I set out to evaluate the difference between narrow and broad peaks.

When calling peaks, generally two types peaks can be obtained according to the cut-off stringency. Narrow peak calling identifies peaks at a higher significance threshold and hence has implications with signal tracks that do not meet certain thresholds. Imputation often times attempts to clean up any noisy signal within 25 base pair bins and dampens the over signal intensity. This can affect the downstream peak calling analysis if the signal track has many peaks and troughs over a short distance. When visualising our signal tracks, I noted that narrow peak calling operates at such a high stringency level, that many peaks become fragmented into smaller regions. This is because the stringency criterion for several bins is not satisfied **(Figure 3.10)**.

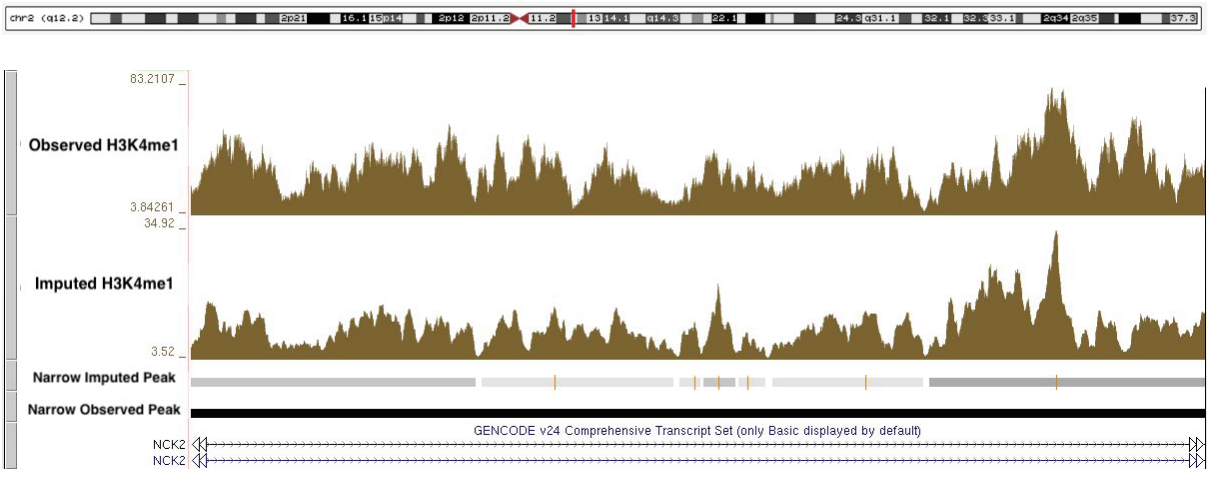


Figure 3.10 Imputation disarranged narrow peaks The sample has been called with narrow peaks, which are the bands below the signal tracks. The imputed peaks have been

broken into many small pieces in comparison to the observed track, and was observed in all the different histone marks. The location is chromosome 2:105,744,897-105,894,274.

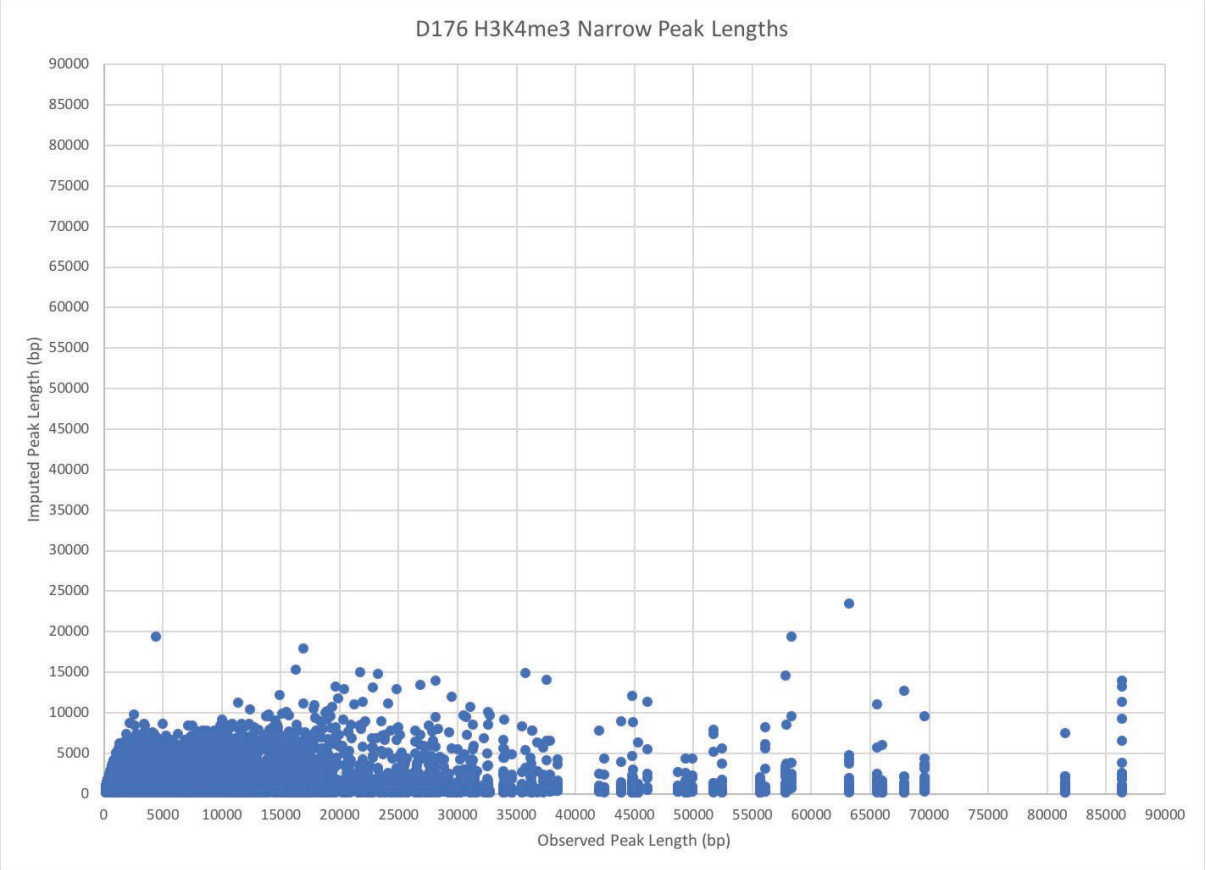


Figure 3.11 Comparing imputed and observed peaks with a narrow peak caller
Narrow peaks were called for both H3K4me3 histone marks for the same sample (D176). After calling, peaks were intersected if the overlap between the peaks was larger than 20 percent. Each of the peaks for the mark would have observed peaks shattered into multiple smaller peaks after the imputation.

Next, I asked if this phenomenon was observed genome wide. I overlapped imputed and observed peaks called with a narrow peak caller. The overlap was done by genomic coordinate. I found that larger observed peaks tended to get fragmented into smaller peaks after imputation, regardless of the histone mark **(Figure 3.11)**.

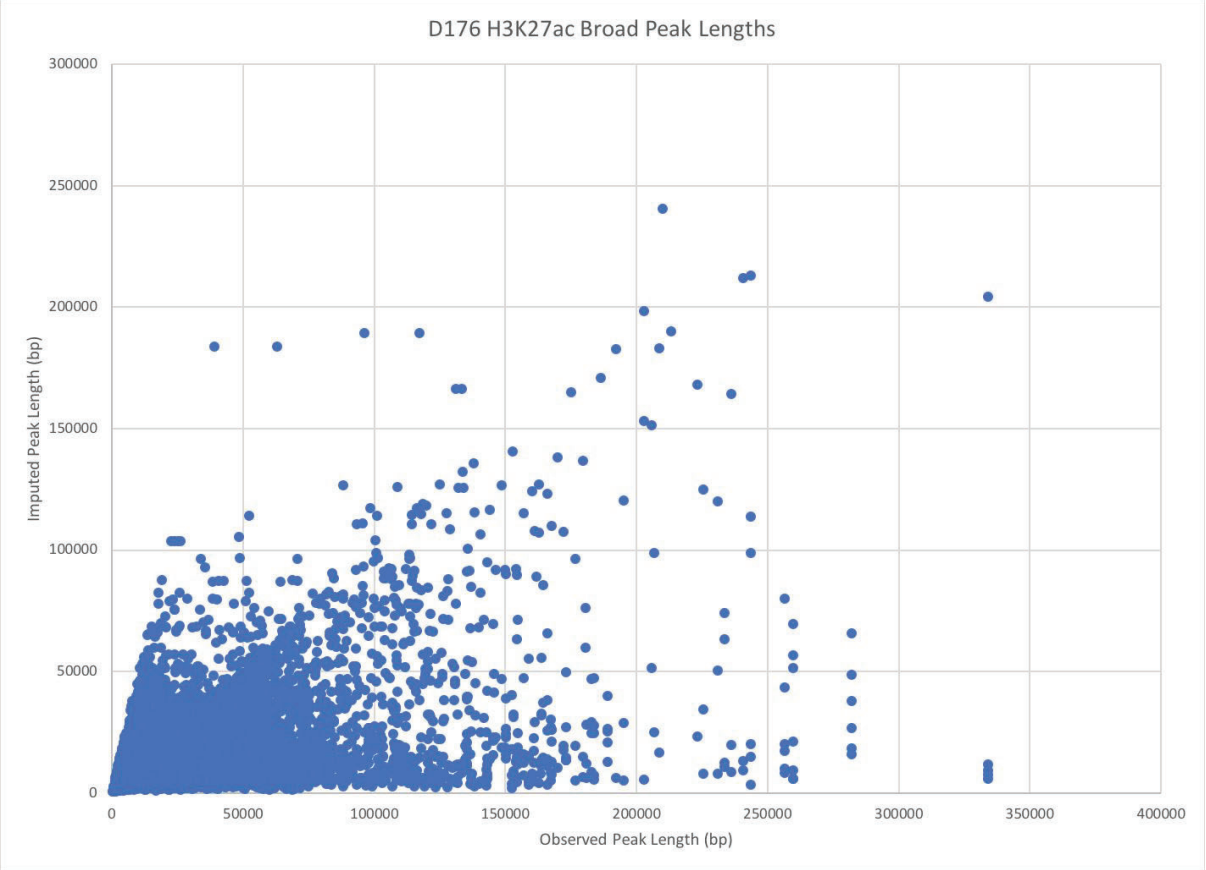


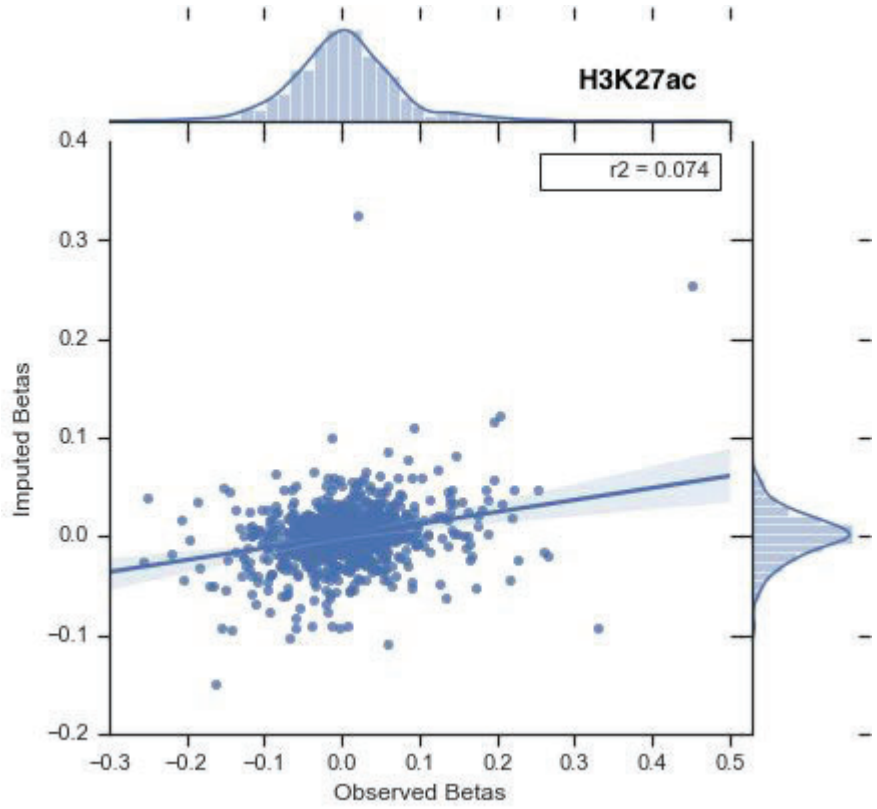
Figure 3.12 Comparing imputed and observed peaks with a broad peak caller
Broad peaks were called for both H3K27ac and histone marks for the same sample (D176). After calling the broad peaks, the imputed peaks would either increase in length or decrease in length. Namely, the shattering seems to disappear when using the broad peak caller.

I then performed the same analysis using a broad peak caller, and intersected the observed and imputed peaks. Peaks obtained in this way seemed to gain length in the imputed data as well as have similar characteristics to the observed peaks **(Figure 3.12)**. It is therefore recommended to use a broad peak caller when analyzing imputed data.

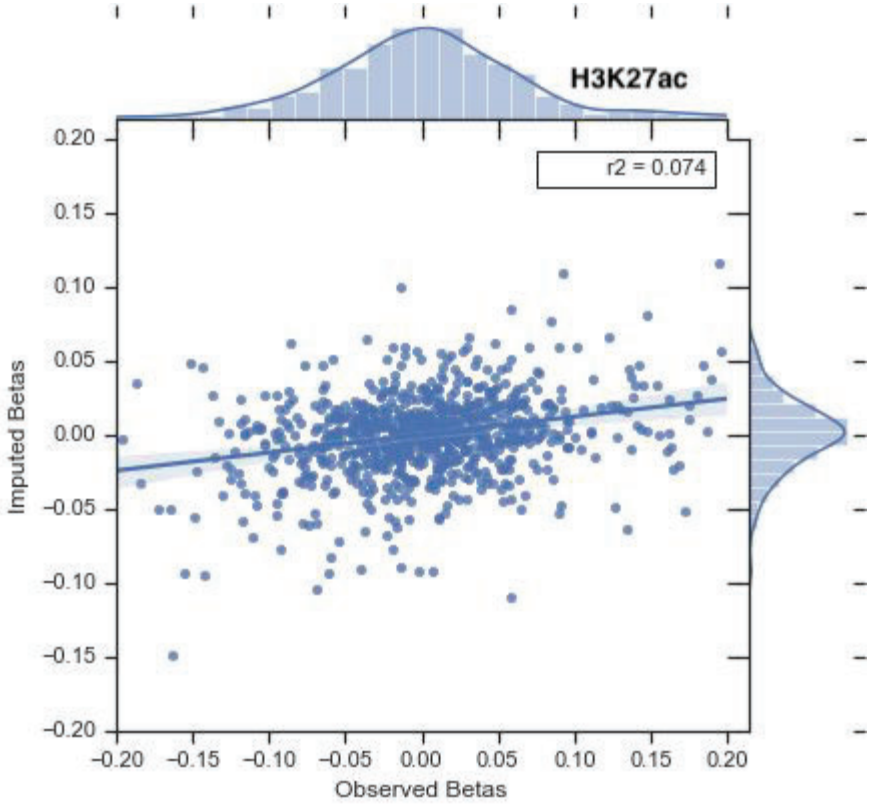
3.6 Imputation disrupts genotypic variability

I was interested to see if ChromImpute would preserve or disrupt genotypic variability in a ChIPmentation data set. I tested both histone QTLs and eQTLs that were initially selected based on their p-values which correspond to the probability of

their effects on gene expression (which are different from 0). I only test those SNPs within a window from the gene (+/- 500kb). Next, I found which eQTLs had a minor allele frequency of at least 30% in our 11 ChIPmentation samples before getting a list of 947 eQTLs. Thus, I picked significant eQTLs that had previously been found as strongly correlated to peaks in a given region. Following the collection of these eQTLs, I then calculated the average signal intensity for those eQTL and the corresponding peaks. The signal intensity was normalized using a Min-Max normalization (Appendix 2), which scaled the signal intensities between 0 and 1. This was done on both observed and imputed tracks for every sample. The values were then associated by genotype value where 0 indicates homozygous dominant, 1 represents heterozygous and 2 represents homozygous recessive. I used linear regression to estimate the effect size (beta) for each eQTL and genotype. The effect sizes were plotted for imputed and observed signal intensities **(Figure 3.13)**. I observed that effect sizes were less variable for the imputed compared to the observed data.



A)



B)

Figure 3.13 Comparison between observed and imputed beta values calculated for eQTLs Investigation of the genetic effects are maintained by the imputation. A selection of 947 peaks that showed strong histone QTL effects (p-value < 0.05) with common minor alleles (minor allele frequency > 0.3). The effects were estimated for Treg samples. Plots here compare the effect sizes of this selection of QTLs for a subset of 11 ChIPmentation samples for the observed and the imputed data. Plot **A)** represents the entire plot while plot **B)** represents the zoomed in plot. The low correlation between the betas indicates that upon imputation the genetic effects are significantly reduced.

There were a few outlier points **(Figure 3.13)** where the signal intensity had varied between the observed and imputed beta's. I picked any points with a beta value larger than 0.2 to visualize the difference in signal intensity between observed and imputed. The signal intensity between the observed and imputed genotypes were fairly random in nature. I observed that many times the imputed signal would either reverse the beta trend or generally lose any genotype specificity, whereas the observed data would have very clear trends and effect sizes **(Figure 3.14)**.

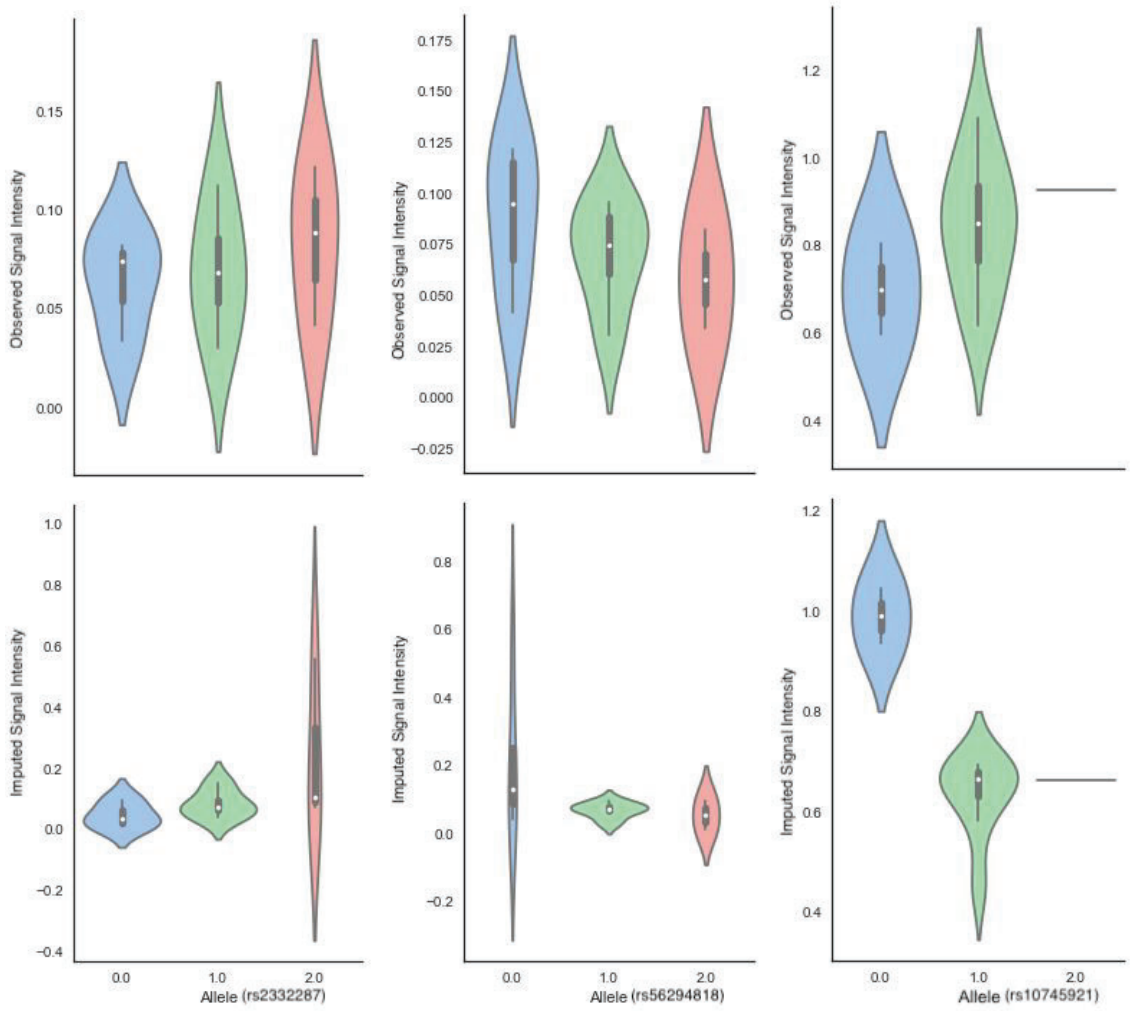


Figure 3.14 eQTL distributions of beta values between observed and imputed signal intensity values Distribution of the signal intensity values from 11 individuals for selected examples of QTL effects after the imputation. In the first example the outlier is not very strong, however the latter examples signify where the beta values significantly differ from the observed. Each of the given violin plots represents the outlier beta from linear regression. Each x-value corresponds to the allele type, with the y-value representing the signal intensity. Imputation does not strengthen any type of genotypic correlation and in fact seems to bear no resemblance to the observed data.

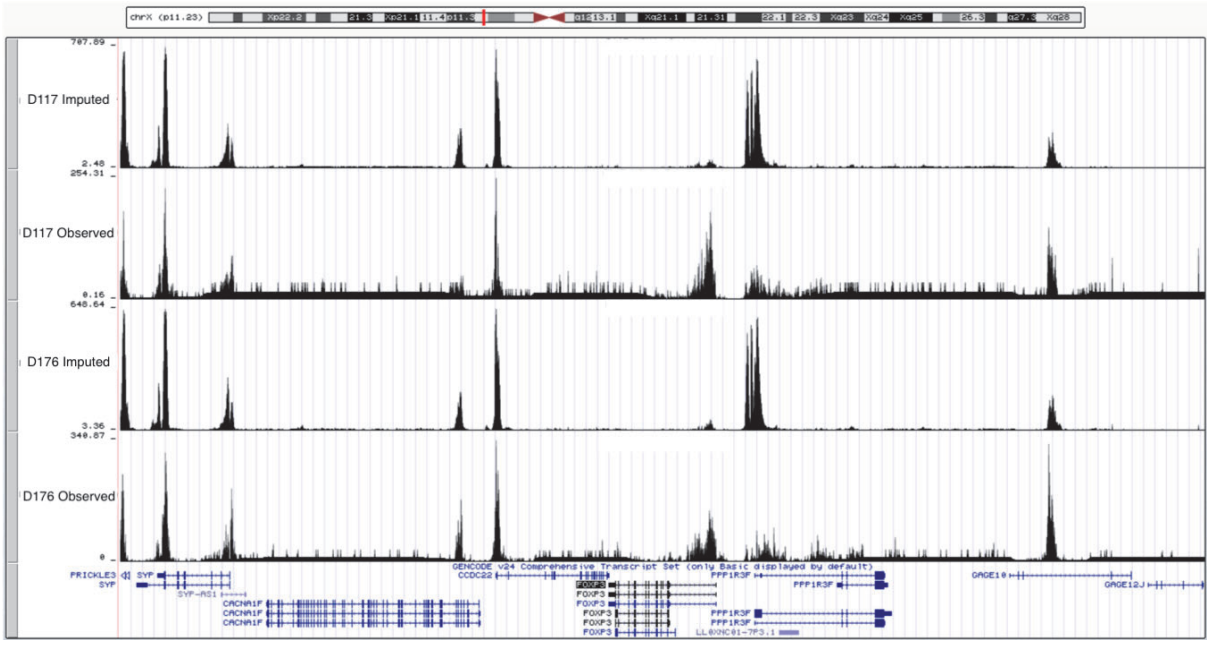
I therefore concluded that ChromImpute does not preserve the genotypic variability, as the genetic effects are often dampened or non-existent.

3.7 Imputation applied to understand Treg biology

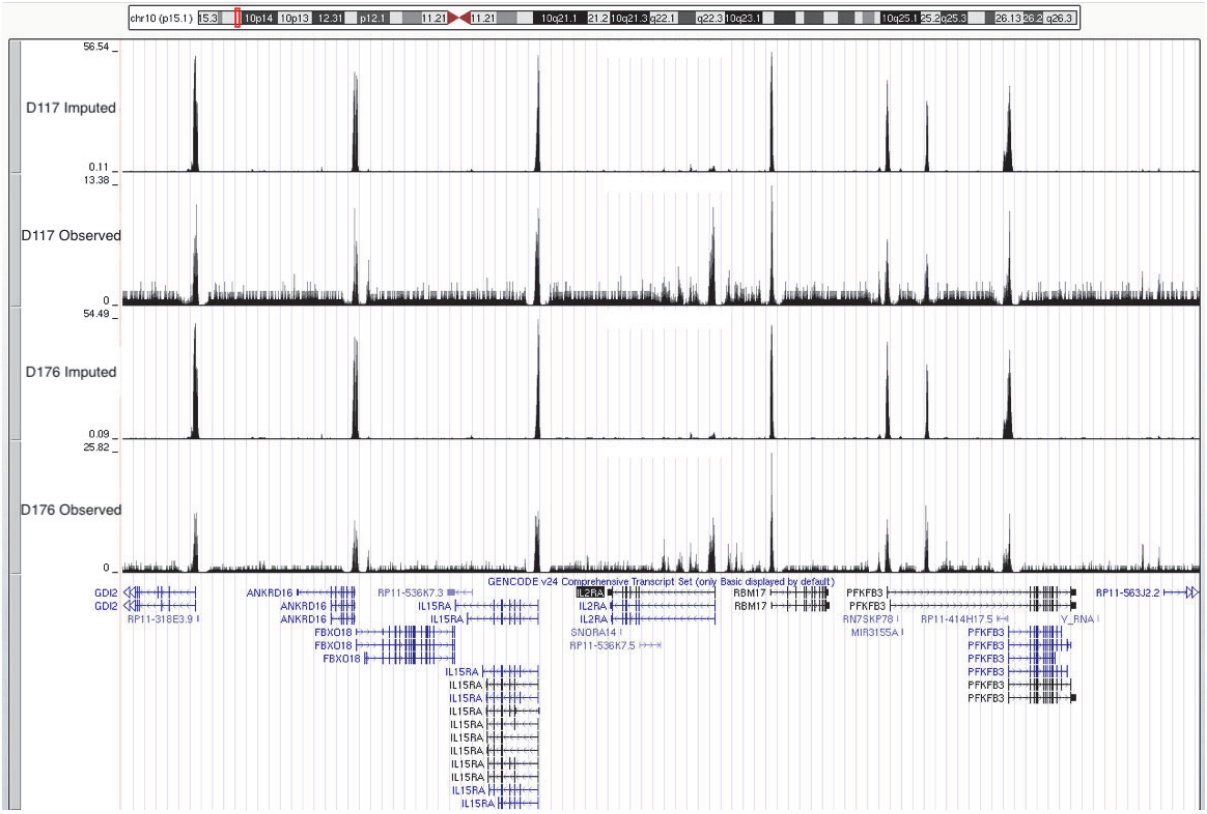
I was interested to see if imputation was able to recapitulate signal specific to Tregs. The challenge of capturing specific signal within a rare cell type is acutely felt, and I wanted to see if imputation amplified overlap for specific Treg genes, as well as any new genes.

I used one of the 3 marks with the best recovery, H3K4me3, to analyze different signal tracks for imputed versus observed for two Treg samples (D117 and D176). I then compared the signal track intensity between the imputed and observed signals where they overlapped a specific Treg gene. I also included a more broad T Cell gene to understand how specific the imputation effects would go, to uncover any specific biological findings.

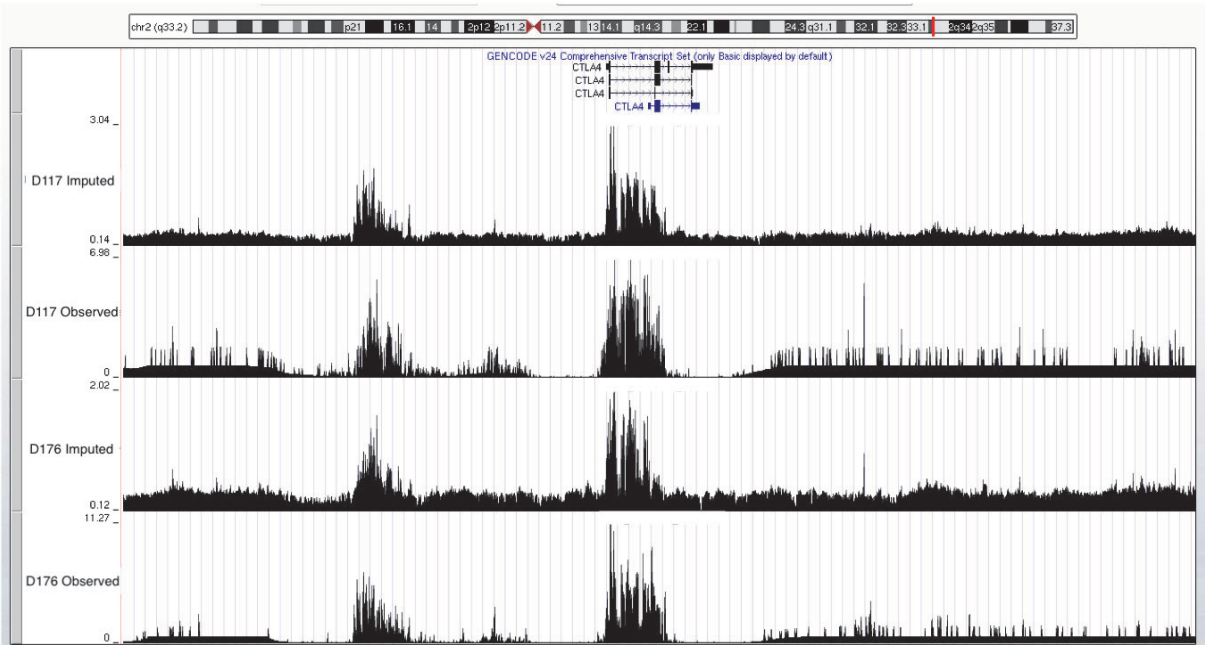
A)



B)


chr10 (p15.1)
D117 Imputed
D117 Observed
D176 Imputed
D176 Observed
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)

C)


chr2 (q33.2)
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
CTLA4
D117 Imputed
D117 Observed
D176 Imputed
D176 Observed

D)

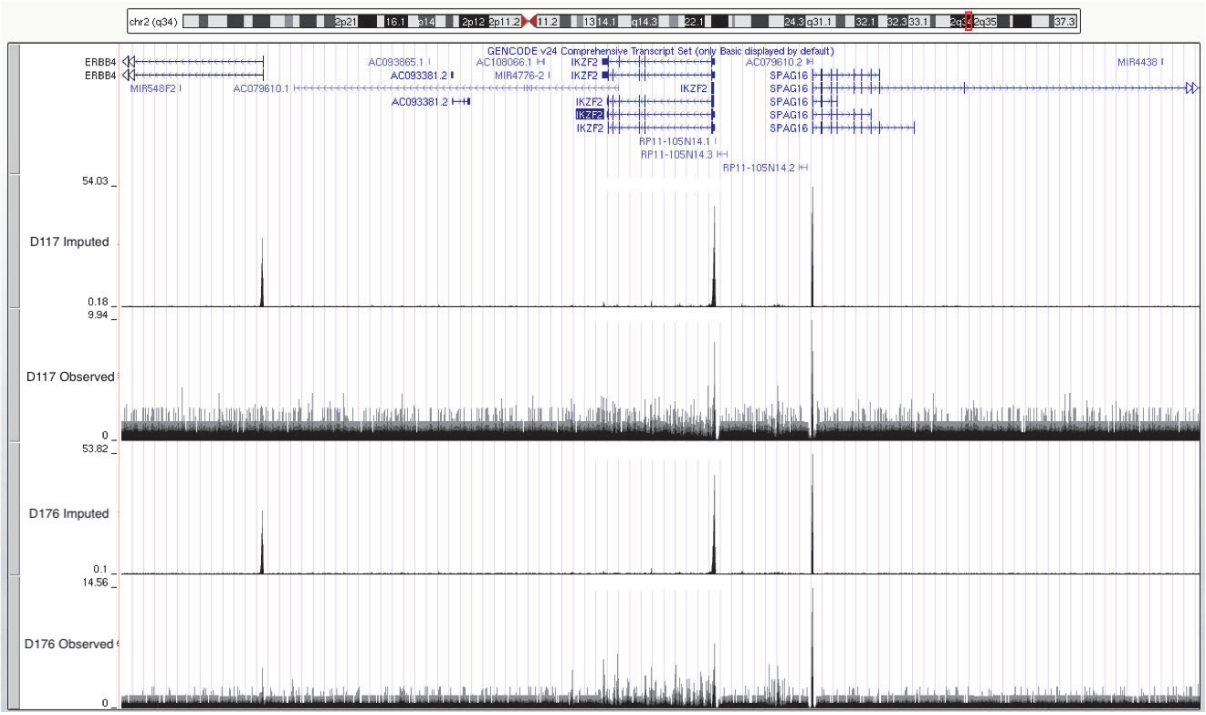


Figure 3.15 Comparison between observed and imputed signal track for key TReg genes Investigation of the key Treg genes are maintained or boosted by the imputation. A selection of 2 different Treg samples for the histone mark H3K4me3. Plots here compare the signal track of for the two samples with observed and imputed. Plot **A)** and plot **B)** evaluate FOXP3 and IL2RA, both highly specific genes to the Treg population, displayed a lot less noise, however lost relative signal with respect to the overlap for the gene. Plot **C)** evaluated another important Treg gene, CTLA4, but had dampened effects after imputation. The final plot **D)** was able to find strong signal in the imputed IKZF2 signal track which codes for a specific transcription factor commonly found in T Cells, zinc finger protein Helios.

I selected 3 key genes specific to Tregs: FOXP3, IL2RA and CTLA4. After plotting the difference peaks in the UCSC genome browser, I noticed that imputation was able to amplify the peaks that were surrounding the area **(Figure 3.15A-C).** However, any overlap with the Treg specific genes were reduced relatively in signal strength. The gene IKZF2, which is a more general T Cell gene, was amplified and the noise surrounding the gene was reduced **(Figure 3.15D)**. I concluded that the

ChromImpute software is able to find signal intensity in broader cell populations but loses a sense of granularity when it comes to rare and specific populations.

4. Discussion and Conclusions

In this thesis, I evaluated ChromImpute applied to multiple datasets and showed the improvements made to histone mark ChIP-seq data as well as potential drawbacks. I provided several benchmarks of ChromImpute against various datasets, and applied ChromImpute to a standard genotypic evaluation. The results suggest that epigenetic imputation improves the quality of epigenetic sequencing information that may be lost from errors during any sequencing steps.

I began by addressing the question of whether or not the global structure of ChM-seq data was preserved after imputation. I showed that this structure is generally maintained when compared to the imputed data as shown by a common set of pathways which are enriched in Treg ChM-seq peaks, and by a reduction of noise when mapping signal to TSS. I then showed that imputation successfully minimizes technical variability, as is evidenced by a reduction in peak variance between observed and imputed peaks. Imputation also corrected for missing signal track in the observed data; this was clearly the case for the CD45 locus (a gene known to be expressed by all Tregs), which recovered its missing signal intensity after imputation. Finally, I found that imputed epigenetic data should generally be analyzed with a broad peak caller in order to provide the best results. This is because imputation provides a very fine-grained signal correction, which causes narrow peaks to be called at every peak and trough, instead of at a global maxima.

One limitation that I explored in this thesis was ChromImpute's ability to account for genotypic variability. ChromImpute generally dampened any differences in intensity observed between individuals. Additionally, when testing for genotype differences and comparing to acetylation QTLs, the directions of effects were fairly random, with some beta's being reversed for no apparent reason. This could be explained by the

ChromImpute software is able to find signal intensity in broader cell populations but loses a sense of granularity when it comes to rare and specific populations.

4. Discussion and Conclusions

In this thesis, I evaluated ChromImpute applied to multiple datasets and showed the improvements made to histone mark ChIP-seq data as well as potential drawbacks. I provided several benchmarks of ChromImpute against various datasets, and applied ChromImpute to a standard genotypic evaluation. The results suggest that epigenetic imputation improves the quality of epigenetic sequencing information that may be lost from errors during any sequencing steps.

I began by addressing the question of whether or not the global structure of ChM-seq data was preserved after imputation. I showed that this structure is generally maintained when compared to the imputed data as shown by a common set of pathways which are enriched in Treg ChM-seq peaks, and by a reduction of noise when mapping signal to TSS. I then showed that imputation successfully minimizes technical variability, as is evidenced by a reduction in peak variance between observed and imputed peaks. Imputation also corrected for missing signal track in the observed data; this was clearly the case for the CD45 locus (a gene known to be expressed by all Tregs), which recovered its missing signal intensity after imputation. Finally, I found that imputed epigenetic data should generally be analyzed with a broad peak caller in order to provide the best results. This is because imputation provides a very fine-grained signal correction, which causes narrow peaks to be called at every peak and trough, instead of at a global maxima.

One limitation that I explored in this thesis was ChromImpute's ability to account for genotypic variability. ChromImpute generally dampened any differences in intensity observed between individuals. Additionally, when testing for genotype differences and comparing to acetylation QTLs, the directions of effects were fairly random, with some beta's being reversed for no apparent reason. This could be explained by the

ChromImpute algorithm only considering signal information from other samples or marks in 25 base pair windows. This would cause any specific genotype effects to be completely expunged during imputation.

Other limitations of ChromImpute concern the bias of the reference panel to the construction of imputed signal tracks. If there are only a few cell types in the reference panel that are closely related to the samples of interest, those cell types will have the biggest influence on the imputed signal track. However, if the reference panel is not diverse in cell types, this can cause samples to lose their inherent features upon imputation. Additionally, if the reference contains samples with a mixture of cell types, the imputed signal tracks will be composed of signal from the same mixture of cells. The deconvolution of these cells is important to maintain the correct signal composition. Imputation depends heavily on the composition of the reference, and this reference can bias the imputed signal tracks. As was uncovered when trying to isolate specific signal intensity for genes with regards to Tregs, ChromImpute was not able to recapitulate that signal. This was due to the reference panel not having enough diversity for Tregs in that given mark. The imputed data was able to capture and isolate signal for a more broad T Cell gene, however. In order for ChromImpute to provide use in this area, the need for incorporating genotype information is a must. These drawbacks to imputation limit its use in population scale genetic studies.

Despite its drawbacks, ChromImpute can be of immense use for analyzing aspects of an experiment which are independent of inter-individual variability, or for overcoming technical biases. As sequencing costs remain high and sample access is scarce, it is important to have tools which help us maximize the quality of data obtained from sequencing experiments. Imputed signal tracks may be useful as a reference catalogue of functional chromatin regions, or as additional samples if needed. For example, if there is a need for any synthetic replicates, ChromImpute can provide an average profile which can increase the overall power of an

experiment. The very nature of the algorithm detects different correlations amongst the samples with different histone marks, which provides the basis of the imputation.

There are several features which should be added in the future to make ChromImpute an integral step in epigenetics data analysis. Firstly, in order to alleviate the problem of low quality samples and reads in the reference biasing the imputation, a quality control (QC) check can be implemented. This control would set a minimum threshold for the number of reads needed for every sample to be part of the reference conglomerate. Secondly, an automatic check on the composition of the reference can be added in order to alert the user of samples that may be over or underrepresented. Lastly, when interpreting the results an automated script can be used to evaluate the accuracy of imputation. This evaluation would be based on he metrics defined in ChromImpute and would rank imputed signal tracks by accuracy.

In order to apply imputation to population scale studies, genotype information should be accounted for. The ability to capture the inter-individual variability within histone marks is extremely difficult to evaluate. The nature of imputation relying on haplotype structure can not be applied here, due to the dynamic nature of chromatin. Chromatin remodeling affects what can be expressed and if one is expected to capture genotypic information, it would be vital to have a conserved structure. This hindrance makes this method to attempt to preserve genotype very difficult.

There are other types of data that can be used to try to find target genes to provide power for these studies. For instance using some type of chromosome conformation capture type data (e.g. 3C or HiC) may provide use if augmented into the method. If the 3D structure of certain non-coding regions is known, that can help add insight into the interactions of a given fragment of a genome. Further, one can then build another point of inference or at the very least eliminate certain possibilities of what the genome enrichment would look like given the given chromatin structure at that given point in time.

Overall the ChromImpute software adds much needed value to any scientists benchside. Epigenetic imputation can be particularly useful in scenarios where experimental assays are costly and time consuming. This tool serves an important purpose and imputed signal tracks can provide a strong reference point and add power to epigenetic studies.

5. Acknowledgements

This year has been one the most enriching I have had, due in large part to the wonderful people I have met. First and foremost, I would like to thank my advisor Dr. Gosia Trynka. Having only lived in the USA for the entirety of my life, I dreamt of being able to live and study out of the country. She single handedly made that dream a reality, and I am incredibly grateful. I appreciate all of her guidance and insight along the way; it is a privilege to work alongside a brilliant and enthusiastic mind like hers. I'd like to additionally thank all the members of the Trynka lab for all their support, guidance and friendship. The amount of time from each of them spent bringing me up to speed and pushing me to finish this project will not be forgotten. I would like to thank my thesis committee: Dr. Leo Parts, Dr. Chris Wallace and Dr. Carl Anderson for their guidance.

I'd like to thank my friends here in the UK, friends back home and family to help me with this transition. Coming to a new country and school after spending time in industry is not the easiest transition and I wouldn't have been able to do it without them.

7. References

Adams, D. et al., 2012. BLUEPRINT to decode the epigenetic signature written in blood. *Nature biotechnology*, 30(3), pp.224–226. Available at: http://dx.doi.org/10.1038/nbt.2153.

Akira, S., Uematsu, S. & Takeuchi, O., 2006. Pathogen recognition and innate immunity. *Cell*, 124(4), pp.783–801. Available at: http://dx.doi.org/10.1016/j.cell.2006.02.015.

Alipanahi, B. et al., 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8), pp.831–838. Available at: http://dx.doi.org/10.1038/nbt.3300.

Anderson, C.A. et al., 2011. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS biology*, 9(1), p.e1000580. Available at: http://dx.doi.org/10.1371/journal.pbio.1000580.

Atabani, S.F. et al., 2005. Association of CTLA4 polymorphism with regulatory T cell frequency. *European journal of immunology*, 35(7), pp.2157–2162. Available at: http://dx.doi.org/10.1002/eji.200526168.

Bannister, A.J. & Kouzarides, T., 2011. Regulation of chromatin by histone modifications. *Cell research*, 21(3), pp.381–395. Available at: http://dx.doi.org/10.1038/cr.2011.22.

Barski, A. et al., 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4), pp.823–837. Available at: http://www.cell.com/article/S0092867407006009/abstract [Accessed August 18, 2018].

Benevolenskaya, E.V., 2007. Histone H3K4 demethylases are essential in development and differentiation. *Biochemistry and cell biology = Biochimie et biologie cellulaire*, 85(4), pp.435–443. Available at: http://dx.doi.org/10.1139/O07-057.

Bernstein, B.E. et al., 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, 28(10), pp.1045–1048. Available at: http://dx.doi.org/10.1038/nbt1010-1045.

Bini, D.A., Latouche, G. & Meini, B., 2005. INTRODUCTION TO MARKOV CHAINS. In *Numerical Methods for Structured Markov Chains*. Oxford: Oxford University Press. Available at:

http://www.oxfordscholarship.com/10.1093/acprof:oso/9780198527688.001.0001/acprof-9780198527688-chapter-1.

Blattler, A. et al., 2014. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome biology*, 15(9), p.469. Available at: http://dx.doi.org/10.1186/s13059-014-0469-0.

Bluestone, J.A., Tang, Q. & Sedwick, C.E., 2008. T regulatory cells in autoimmune diabetes: past challenges, future prospects. *Journal of clinical immunology*, 28(6), pp.677–684. Available at: http://dx.doi.org/10.1007/s10875-008-9242-z.

Bock, C. & Lengauer, T., 2008. Computational epigenetics. *Bioinformatics* , 24(1), pp.1–10. Available at: http://dx.doi.org/10.1093/bioinformatics/btm546.

Browning, S.R. & Browning, B.L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics*, 81(5), pp.1084–1097. Available at: http://dx.doi.org/10.1086/521987.

Buenrostro, J.D. et al., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, 10, p.1213. Available at: http://dx.doi.org/10.1038/nmeth.2688.

Bush, W.S. & Moore, J.H., 2012. Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12), p.e1002822. Available at: http://dx.doi.org/10.1371/journal.pcbi.1002822.

Cantor, R.M., Lange, K. & Sinsheimer, J.S., 2010. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *American journal of human genetics*, 86(1), pp.6–22. Available at: http://dx.doi.org/10.1016/j.ajhg.2009.11.017.

Chen, Y. et al., 2012. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature methods*, 9(6), pp.609–614. Available at: http://dx.doi.org/10.1038/nmeth.1985.

Cock, P.J.A. et al., 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6), pp.1767–1771. Available at: http://dx.doi.org/10.1093/nar/gkp1137.

Colombel, J.-F. et al., 2007. Adalimumab for maintenance of clinical response and remission in patients with Crohn's disease: the CHARM trial. *Gastroenterology*, 132(1), pp.52–65. Available at: http://dx.doi.org/10.1053/j.gastro.2006.11.041.

Dale, R.K., Pedersen, B.S. & Quinlan, A.R., 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* , 27(24), pp.3423–3424. Available at:

http://dx.doi.org/10.1093/bioinformatics/btr539.

Das, R. et al., 2006. Computational prediction of methylation status in human genomic sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 103(28), pp.10713–10716. Available at: http://dx.doi.org/10.1073/pnas.0602949103.

Eddy, S.R., 1998. Profile hidden Markov models. *Bioinformatics* , 14(9), pp.755–763. Available at: https://www.ncbi.nlm.nih.gov/pubmed/9918945.

Edwards, S.L. et al., 2013. Beyond GWASs: illuminating the dark road from association to function. *American journal of human genetics*, 93(5), pp.779–797. Available at: http://dx.doi.org/10.1016/j.ajhg.2013.10.012.

Encinas, J.A. et al., 1999. QTL influencing autoimmune diabetes and encephalomyelitis map to a 0.15-cM region containing Il2. *Nature genetics*, 21(2), pp.158–160. Available at: http://dx.doi.org/10.1038/5941.

ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74. Available at: http://dx.doi.org/10.1038/nature11247.

ENCODE Project Consortium et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), pp.799–816. Available at: http://dx.doi.org/10.1038/nature05874.

ENCODE Project Consortium, 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696), pp.636–640. Available at: http://dx.doi.org/10.1126/science.1105136.

Ernst, J. & Kellis, M., 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, 28(8), pp.817–825. Available at: http://dx.doi.org/10.1038/nbt.1662.

Ernst, J. & Kellis, M., 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature biotechnology*, 33, p.364. Available at: http://dx.doi.org/10.1038/nbt.3157.

Evans, J.P. et al., 2011. Genomics. Deflating the genomic bubble. *Science*, 331(6019), pp.861–862. Available at: http://dx.doi.org/10.1126/science.1198039.

Fletcher, J.M. et al., 2009. CD39+Foxp3+ regulatory T Cells suppress pathogenic Th17 cells and are impaired in multiple sclerosis. *Journal of immunology* , 183(11), pp.7602–7610. Available at: http://dx.doi.org/10.4049/jimmunol.0901881.

Freedman, M.L. et al., 2011. Principles for the post-GWAS functional characterization of cancer risk loci. *Nature genetics*, 43(6), pp.513–518.

Available at: http://dx.doi.org/10.1038/ng.840.

Furst, D.E., 2010. The risk of infections with biologic therapies for rheumatoid arthritis. *Seminars in arthritis and rheumatism*, 39(5), pp.327–346. Available at: http://dx.doi.org/10.1016/j.semarthrit.2008.10.002.

Gabriel, S.B. et al., 2002. The structure of haplotype blocks in the human genome. *Science*, 296(5576), pp.2225–2229. Available at: http://dx.doi.org/10.1126/science.1069424.

Gates, L.A. et al., 2017. Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. *The Journal of biological chemistry*. Available at: http://www.jbc.org/content/early/2017/07/17/jbc.M117.802074.abstract.

Guerini, F.R. et al., 2012. A functional variant in ERAP1 predisposes to multiple sclerosis. *PloS one*, 7(1), p.e29931. Available at: http://dx.doi.org/10.1371/journal.pone.0029931.

Hoffmann, T.J. & Witte, J.S., 2015. Strategies for Imputing and Analyzing Rare Variants in Association Studies. *Trends in genetics: TIG*, 31(10), pp.556–563. Available at: http://dx.doi.org/10.1016/j.tig.2015.07.006.

Holland, S.M. et al., 2007. STAT3 mutations in the hyper-IgE syndrome. *The New England journal of medicine*, 357(16), pp.1608–1619. Available at: http://dx.doi.org/10.1056/NEJMoa073687.

Howie, B. et al., 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, 44(8), pp.955–959. Available at: http://dx.doi.org/10.1038/ng.2354.

Howie, B.N., Donnelly, P. & Marchini, J., 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6), p.e1000529. Available at: http://dx.doi.org/10.1371/journal.pgen.1000529.

Hrdlickova, B. et al., 2014. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochimica et biophysica acta*, 1842(10), pp.1910–1922. Available at: http://dx.doi.org/10.1016/j.bbadis.2014.03.011.

Hsu, F. et al., 2006. The UCSC Known Genes. *Bioinformatics* , 22(9), pp.1036–1046. Available at: https://academic.oup.com/bioinformatics/article/22/9/1036/200093 [Accessed August 1, 2018].

Hunter, D.J. & Kraft, P., 2007. Drinking from the fire hose--statistical issues in genomewide association studies. *The New England journal of medicine*, 357(5),

pp.436–439. Available at: http://dx.doi.org/10.1056/NEJMp078120.

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, 9(3), pp.90–95. Available at: http://dx.doi.org/10.1109/MCSE.2007.55.

Jerez, J.M. et al., 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, 50(2), pp.105–115. Available at: http://dx.doi.org/10.1016/j.artmed.2010.05.002.

Jung, Y.L. et al., 2014. Impact of sequencing depth in ChIP-seq experiments. *Nucleic acids research*, 42(9), p.e74. Available at: http://dx.doi.org/10.1093/nar/gku178.

Karlić, R. et al., 2010. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7), pp.2926–2931. Available at: http://www.pnas.org/content/107/7/2926 [Accessed August 21, 2018].

Kent, W.J. et al., 2002. The human genome browser at UCSC. *Genome research*, 12(6), pp.996–1006. Available at: http://dx.doi.org/10.1101/gr.229102.

Kluyver, T. et al., 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Scmidt, eds. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press, pp. 87–90. Available at: https://eprints.soton.ac.uk/403913/ [Accessed August 1, 2018].

Korber, P. et al., 2004. Evidence for histone eviction in trans upon induction of the yeast PHO5 promoter. *Molecular and cellular biology*, 24(24), pp.10965–10974. Available at: http://dx.doi.org/10.1128/MCB.24.24.10965-10974.2004.

Korte, A. & Farlow, A., 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, 9, p.29. Available at: http://dx.doi.org/10.1186/1746-4811-9-29.

Kouzarides, T., 2007. Chromatin modifications and their function. *Cell*, 128(4), pp.693–705. Available at: http://dx.doi.org/10.1016/j.cell.2007.02.005.

Kriegler, M. et al., 1988. A novel form of TNF/cachectin is a cell surface cytotoxic transmembrane protein: ramifications for the complex physiology of TNF. *Cell*, 53(1), pp.45–53. Available at: https://www.ncbi.nlm.nih.gov/pubmed/3349526.

Krogh, A., 1998. An Introduction to Hidden Markov Models for Biological Sequences. In D. B. S. A. S. K. S. L. Salzberg, ed. *Computational Methods in Molecular Biology*. Elsevier, pp. 45–63. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.7972&rep=rep1&ty

pe=pdf.

Kronenberg, M. & Rudensky, A., 2005. Regulation of immunity by self-reactive T cells. *Nature*, 435(7042), pp.598–604. Available at: http://dx.doi.org/10.1038/nature03725.

Landt, S.G. et al., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9), pp.1813–1831. Available at: http://dx.doi.org/10.1101/gr.136184.111.

Lauberth, S.M. et al., 2013. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell*, 152(5), pp.1021–1036. Available at: http://dx.doi.org/10.1016/j.cell.2013.01.052.

Liang, G. et al., 2004. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19), pp.7357–7362. Available at: http://dx.doi.org/10.1073/pnas.0401866101.

Librado, P. & Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* , 25(11), pp.1451–1452. Available at: http://dx.doi.org/10.1093/bioinformatics/btp187.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*. Available at: http://arxiv.org/abs/1303.3997.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , 25(16), pp.2078–2079. Available at: http://dx.doi.org/10.1093/bioinformatics/btp352.

Liu, J. et al., 2015. Chromatin Landscape Defined by Repressive Histone Methylation during Oligodendrocyte Differentiation. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 35(1), pp.352–365. Available at: http://www.jneurosci.org/content/35/1/352 [Accessed August 23, 2018].

Li, X. & Zheng, Y., 2015. Regulatory T cell identity: formation and maintenance. *Trends in immunology*, 36(6), pp.344–353. Available at: http://dx.doi.org/10.1016/j.it.2015.04.006.

Li, Y. et al., 2009. Genotype imputation. *Annual review of genomics and human genetics*, 10, pp.387–406. Available at: http://dx.doi.org/10.1146/annurev.genom.9.081307.164242.

Lorch, Y., Maier-Davis, B. & Kornberg, R.D., 2010. Mechanism of chromatin remodeling. *Proceedings of the National Academy of Sciences of the United States of America*, 107(8), pp.3458–3462. Available at:

http://dx.doi.org/10.1073/pnas.1000398107.

Manolio, T.A., 2013. Bringing genome-wide association findings into clinical use. *Nature reviews. Genetics*, 14, p.549. Available at: http://dx.doi.org/10.1038/nrg3523.

Manolio, T.A. et al., 2009. Finding the missing heritability of complex diseases. *Nature*, 461(7265), pp.747–753. Available at: http://dx.doi.org/10.1038/nature08494.

Marchini, J. & Howie, B., 2010. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics*, 11(7), pp.499–511. Available at: http://dx.doi.org/10.1038/nrg2796.

Martin, E.R. et al., 2000. SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *American journal of human genetics*, 67(2), pp.383–394. Available at: http://dx.doi.org/10.1086/303003.

McCarthy, S. et al., 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10), pp.1279–1283. Available at: http://dx.doi.org/10.1038/ng.3643.

McKerns, M.M. et al., 2012. Building a Framework for Predictive Science. *arXiv [cs.MS]*. Available at: http://arxiv.org/abs/1202.1056.

Mc Kinney, W., pandas: a Foundational Python Library for Data Analysis and Statistics. Available at: https://www.dlr.de/sc/Portaldata/15/Resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf.

McLean, C.Y. et al., 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5), pp.495–501. Available at: http://dx.doi.org/10.1038/nbt.1630.

Mertes, F. et al., 2011. Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings in functional genomics*, 10(6), pp.374–386. Available at: https://academic.oup.com/bfg/article/10/6/374/233492 [Accessed August 20, 2018].

Mikkelsen, T.S. et al., 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153), pp.553–560. Available at: http://dx.doi.org/10.1038/nature06008.

Millman, K.J. & Aivazis, M., 2011. Python for Scientists and Engineers. *Computing in Science Engineering*, 13(2), pp.9–12. Available at: http://dx.doi.org/10.1109/MCSE.2011.36.

NCI-NHGRI Working Group on Replication in Association Studies et al., 2007. Replicating genotype-phenotype associations. *Nature*, 447(7145), pp.655–660. Available at: http://dx.doi.org/10.1038/447655a.

O'Geen, H., Echipare, L. & Farnham, P.J., 2011. Using ChIP-seq technology to generate high-resolution profiles of histone modifications. *Methods in molecular biology* , 791, pp.265–286. Available at: http://dx.doi.org/10.1007/978-1-61779-316-5_20.

Oliphant, T.E., 2006. *A guide to NumPy*, Trelgol Publishing USA. Available at: http://web.mit.edu/dvp/Public/numpybook.pdf.

Oliphant, T.E., 2007. Python for Scientific Computing. *Computing in Science Engineering*, 9(3), pp.10–20. Available at: http://dx.doi.org/10.1109/MCSE.2007.58.

Park, P.J., 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10), pp.669–680. Available at: http://dx.doi.org/10.1038/nrg2641.

Pedregosa, F. et al., 2011. Scikit-learn: Machine Learning in Python. *Journal of machine learning research: JMLR*, 12(Oct), pp.2825–2830. Available at: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf [Accessed August 1, 2018].

Pei, Y.-F. et al., 2010. Analyses and comparison of imputation-based association methods. *PloS one*, 5(5), p.e10827. Available at: http://dx.doi.org/10.1371/journal.pone.0010827.

Perez, F. & Granger, B.E., 2007. IPython: A System for Interactive Scientific Computing. *Computing in Science Engineering*, 9(3), pp.21–29. Available at: http://dx.doi.org/10.1109/MCSE.2007.53.

Pidasheva, S. et al., 2011. Functional studies on the IBD susceptibility gene IL23R implicate reduced receptor function in the protective genetic variant R381Q. *PloS one*, 6(10), p.e25038. Available at: http://dx.doi.org/10.1371/journal.pone.0025038.

Pierini, A. et al., 2014. CD4+FoxP3+ Regulatory T Cells Regulate B Cell Differentiation and Induce Tolerance to Bone Marrow Grafts. *Blood*, 124(21), pp.4322–4322. Available at: http://www.bloodjournal.org/content/124/21/4322?sso-checked=true [Accessed August 19, 2018].

Portales-Casamar, E. et al., 2010. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic acids research*, 38(Database issue), pp.D105–10. Available at:

http://dx.doi.org/10.1093/nar/gkp950.

Purcell, S. et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3), pp.559–575. Available at: http://dx.doi.org/10.1086/519795.

Quail, M.A. et al., 2008. A large genome center's improvements to the Illumina sequencing system. *Nature methods*, 5(12), pp.1005–1010. Available at: http://dx.doi.org/10.1038/nmeth.1270.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* , 26(6), pp.841–842. Available at: http://dx.doi.org/10.1093/bioinformatics/btq033.

Radman-Livaja, M. & Rando, O.J., 2010. Nucleosome positioning: how is it established, and why does it matter? *Developmental biology*, 339(2), pp.258–266. Available at: http://dx.doi.org/10.1016/j.ydbio.2009.06.012.

Ramos, R.G. & Olden, K., 2008. Gene-environment interactions in the development of complex disease phenotypes. *International journal of environmental research and public health*, 5(1), pp.4–11. Available at: https://www.ncbi.nlm.nih.gov/pubmed/18441400.

Raychaudhuri, S. et al., 2012. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nature genetics*, 44(3), pp.291–296. Available at: http://dx.doi.org/10.1038/ng.1076.

Raychaudhuri, S., 2011. Mapping rare and common causal alleles for complex human diseases. *Cell*, 147(1), pp.57–69. Available at: http://dx.doi.org/10.1016/j.cell.2011.09.011.

Rea, S. et al., 2000. Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature*, 406(6796), pp.593–599. Available at: http://dx.doi.org/10.1038/35020506.

Rioux, J.D. & Abbas, A.K., 2005. Paths to understanding the genetic basis of autoimmune disease. *Nature*, 435(7042), pp.584–589. Available at: http://dx.doi.org/10.1038/nature03723.

Rivera, C.M. & Ren, B., 2013. Mapping Human Epigenomes. *Cell*, 155(1), pp.39–55. Available at: https://www.cell.com/fulltext/S0092-8674(13)01148-3 [Accessed August 21, 2018].

Robertson, G. et al., 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8), pp.651–657. Available at: http://dx.doi.org/10.1038/nmeth1068.

Rosenblum, M.D., Remedios, K.A. & Abbas, A.K., 2015. Mechanisms of human

autoimmunity. *The Journal of clinical investigation*, 125(6), pp.2228–2233. Available at: http://dx.doi.org/10.1172/JCI78088.

Rosenfeld, J.A. et al., 2009. Determination of enriched histone modifications in non-genic portions of the human genome. *BMC genomics*, 10, p.143. Available at: http://dx.doi.org/10.1186/1471-2164-10-143.

Roshyara, N.R. et al., 2016. Comparing performance of modern genotype imputation methods in different ethnicities. *Scientific reports*, 6, p.34386. Available at: http://dx.doi.org/10.1038/srep34386.

Rutgeerts, P. et al., 2005. Infliximab for induction and maintenance therapy for ulcerative colitis. *The New England journal of medicine*, 353(23), pp.2462–2476. Available at: http://dx.doi.org/10.1056/NEJMoa050516.

Sakaguchi, S. et al., 1995. Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor alpha-chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. *Journal of immunology* , 155(3), pp.1151–1164. Available at: https://www.ncbi.nlm.nih.gov/pubmed/7636184.

Sakaguchi, S. et al., 2008. Regulatory T cells and immune tolerance. *Cell*, 133(5), pp.775–787. Available at: http://dx.doi.org/10.1016/j.cell.2008.05.009.

Schmidl, C. et al., 2015. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nature methods*, 12(10), pp.963–965. Available at: http://dx.doi.org/10.1038/nmeth.3542.

Schones, D.E. et al., 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5), pp.887–898. Available at: http://dx.doi.org/10.1016/j.cell.2008.02.022.

Schork, A.J. et al., 2013. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS genetics*, 9(4), p.e1003449. Available at: http://dx.doi.org/10.1371/journal.pgen.1003449.

Shu, W. et al., 2011. Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. *Nucleic acids research*, 39(17), pp.7428–7443. Available at: http://dx.doi.org/10.1093/nar/gkr443.

Spain, S.L. & Barrett, J.C., 2015. Strategies for fine-mapping complex traits. *Human molecular genetics*, 24(R1), pp.R111–9. Available at: http://dx.doi.org/10.1093/hmg/ddv260.

Spencer, C.C.A. et al., 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS genetics*, 5(5),

p.e1000477. Available at: http://dx.doi.org/10.1371/journal.pgen.1000477.

Starmer, J. & Magnuson, T., 2016. Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains. *BMC bioinformatics*, 17, p.144. Available at: http://dx.doi.org/10.1186/s12859-016-0991-z.

Stunnenberg, H.G. et al., 2016. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, 167(5), pp.1145–1149. Available at: https://www.cell.com/cell/abstract/S0092-8674(16)31528-8 [Accessed August 18, 2018].

Tehranchi, A.K. et al., 2016. Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell*, 165(3), pp.730–741. Available at: http://dx.doi.org/10.1016/j.cell.2016.03.041.

The International HapMap 3 Consortium et al., 2010. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467, p.52. Available at: http://dx.doi.org/10.1038/nature09298.

Thurman, R.E. et al., 2012. The accessible chromatin landscape of the human genome. *Nature*, 489(7414), pp.75–82. Available at: http://dx.doi.org/10.1038/nature11232.

Troyanskaya, O. et al., 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* , 17(6), pp.520–525. Available at: https://www.ncbi.nlm.nih.gov/pubmed/11395428.

Trynka, G. et al., 2013. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics*, 45(2), pp.124–130. Available at: http://dx.doi.org/10.1038/ng.2504.

Trynka, G. & Raychaudhuri, S., 2013. Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases. *Current opinion in genetics & development*, 23(6), pp.635–641. Available at: http://dx.doi.org/10.1016/j.gde.2013.10.009.

Tsompana, M. & Buck, M.J., 2014. Chromatin accessibility: a window into the genome. *Epigenetics & chromatin*, 7(1), p.33. Available at: https://doi.org/10.1186/1756-8935-7-33.

Valouev, A. et al., 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods*, 5(9), pp.829–834. Available at: http://dx.doi.org/10.1038/nmeth.1246.

Varmus, H., 2010. Ten years on--the human genome and medicine. *The New England journal of medicine*, 362(21), pp.2028–2029. Available at: http://dx.doi.org/10.1056/NEJMe0911933.

Visel, A. et al., 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231), pp.854–858. Available at: http://dx.doi.org/10.1038/nature07730.

Visscher, P.M. et al., 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *American journal of human genetics*, 101(1), pp.5–22. Available at: http://dx.doi.org/10.1016/j.ajhg.2017.06.005.

Wang, L. et al., 2014. Breach of tolerance: primary biliary cirrhosis. *Seminars in liver disease*, 34(3), pp.297–317. Available at: http://dx.doi.org/10.1055/s-0034-1383729.

Waskom, M. et al., 2014. *seaborn: v0.5.0 (November 2014)*, Available at: https://zenodo.org/record/12710.

Wellcome Trust Case Control Consortium et al., 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289), pp.713–720. Available at: http://dx.doi.org/10.1038/nature08979.

Yu, G., Wang, L.-G. & He, Q.-Y., 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* , 31(14), pp.2382–2383. Available at: http://dx.doi.org/10.1093/bioinformatics/btv145.

Zhang, F. & Lupski, J.R., 2015. Non-coding genetic variants in human disease. *Human molecular genetics*, 24(R1), pp.R102–10. Available at: http://dx.doi.org/10.1093/hmg/ddv259.

Zhang, Y. et al., 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome biology*, 9(9), p.R137. Available at: https://doi.org/10.1186/gb-2008-9-9-r137.

Zhang, Z. & Pugh, B.F., 2011. High-resolution genome-wide mapping of the primary structure of chromatin. *Cell*, 144(2), pp.175–186. Available at: http://dx.doi.org/10.1016/j.cell.2011.01.003.

Zhou, J. et al., 2011. A20-binding inhibitor of NF-κB (ABIN1) controls Toll-like receptor-mediated CCAAT/enhancer-binding protein β activation and protects from inflammatory disease. *Proceedings of the National Academy of Sciences of the United States of America*, 108(44), pp.E998–1006. Available at: http://dx.doi.org/10.1073/pnas.1106232108.

Zhou, J. & Troyanskaya, O.G., 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*, 12(10), pp.931–934. Available at: http://dx.doi.org/10.1038/nmeth.3547.

6. Abbreviation List

Treg	Regulatory T Cell
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
TNFα	Tumor necrosis factor - alpha
GWAS	Genome-wide association studies
SNPs	Single nucleotide polymorphisms
LD	Linkage disequilibrium
ATAC-seq	Transposase-accessible chromatin followed by sequencing
HMM	Hidden markov models
ChM-seq	ChIPmentation-seq
GRCh38	Human reference genome build 38
BED	Browser extensible data
MACS	Model-based analysis of chIP-seq
ENCODE	Encyclopedia of DNA elements
3C	Chromosome conformation capture
BAM	Binary alignment map
ROC	Receiver operating characteristic

RAM	**Random access memory**
UCSC	**University of California - Santa Cruz**
TSS	**Transcription start sites**
KEGG	**Kyoto encyclopedia of genes and genomes**
UPGMA	**Unweighted pair group method with arithmetic mean**
PTPRC	**Protein tyrosine phosphatase, receptor type, C**
QTL	**Quantitative trait locus**
QC	**Quality control**