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).

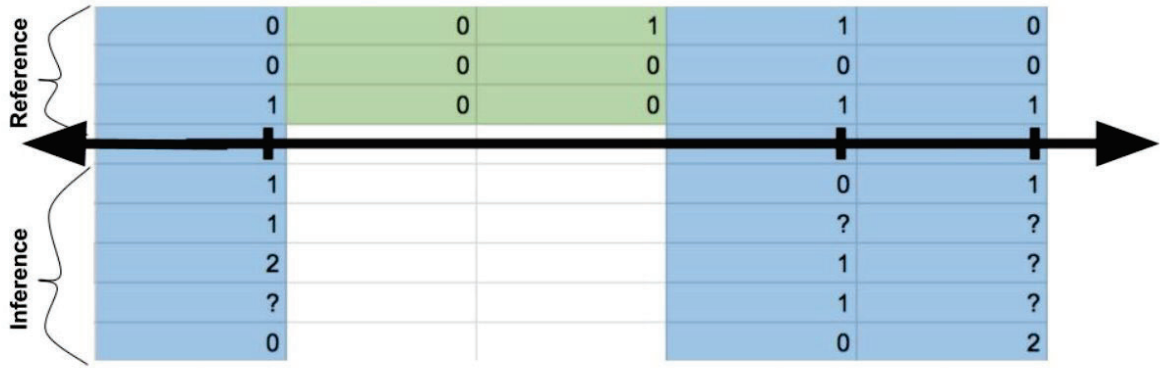

Figure 1.2 Imputation in GWAS General procedure for imputing genotype information. First, samples are phased at the genotyped SNPs. Next, SNPs are compared to the reference haplotypes (ex. HapMap 3 (The International HapMap 3 Consortium et al. 2010)). Sample haplotypes are matched to the best mixture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

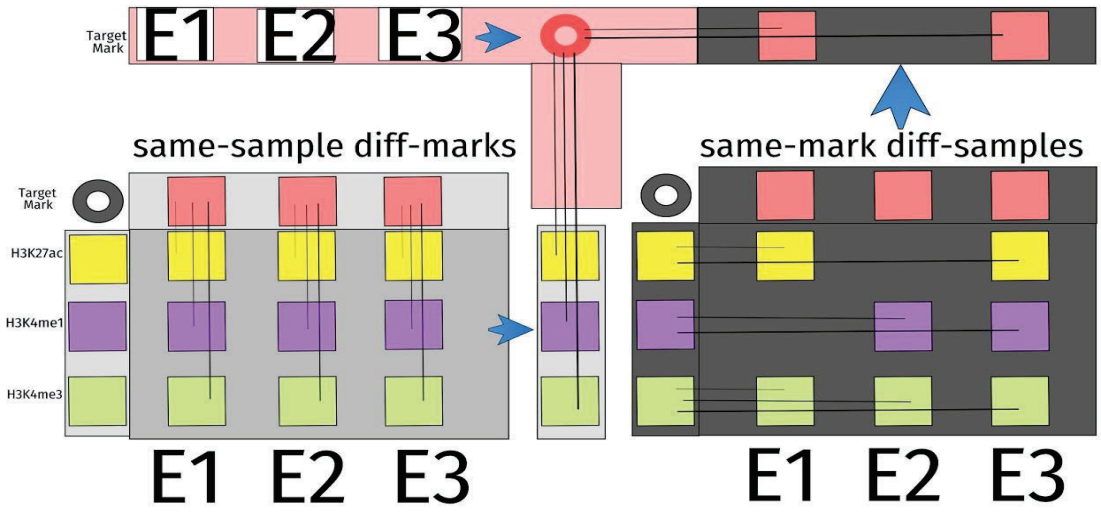

Figure 1.3 ChromImpute method This software begins by taking different samples (e.g. E1, E2, E3) and tabulating each of them based on the different marks within a given sample and then finding other samples with the same mark. Once that process has finished regression trees are built around the different 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