3. Results

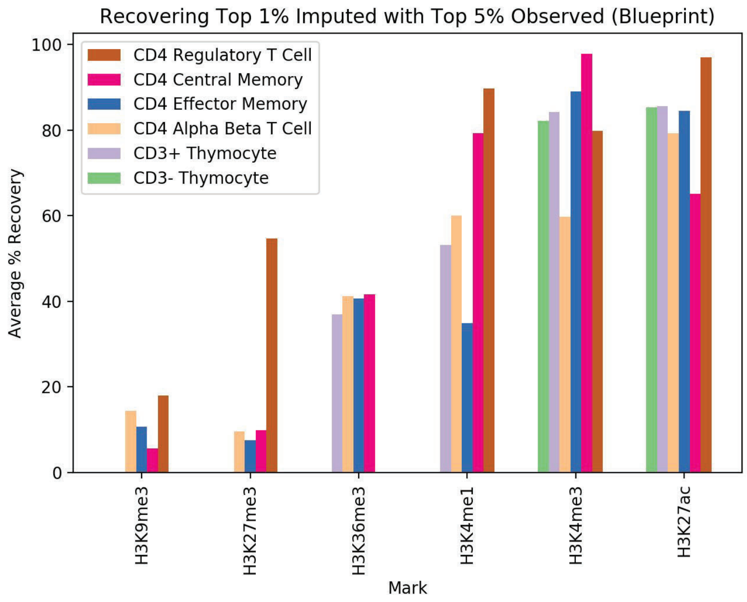
3.1 Overview

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

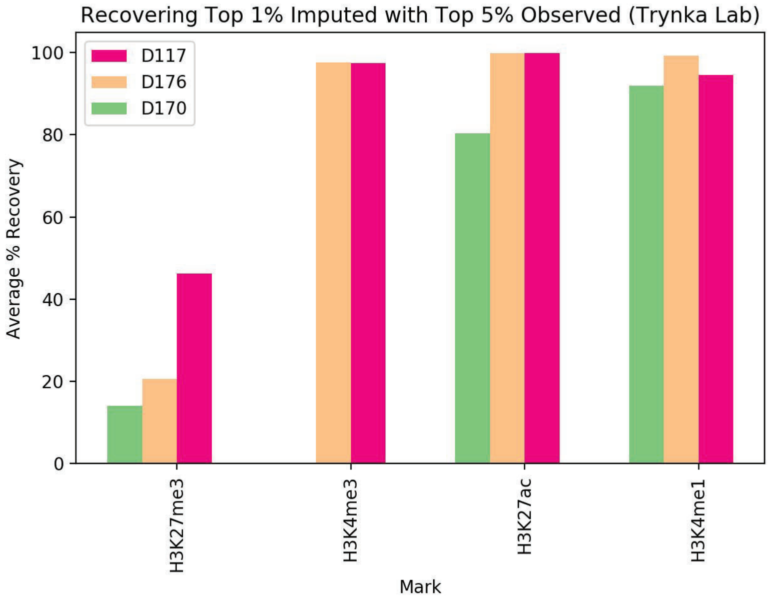


Figure 3.1 Imputation accuracy This metric evaluates the top one percent of signal bins in the imputed data that is also in the observed data. The Blueprint samples, **A)** showed the highest recovery in H3K27ac and H3K4me3 marks reads used in the reference. Similarly, the Trynka lab samples **B)** showed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

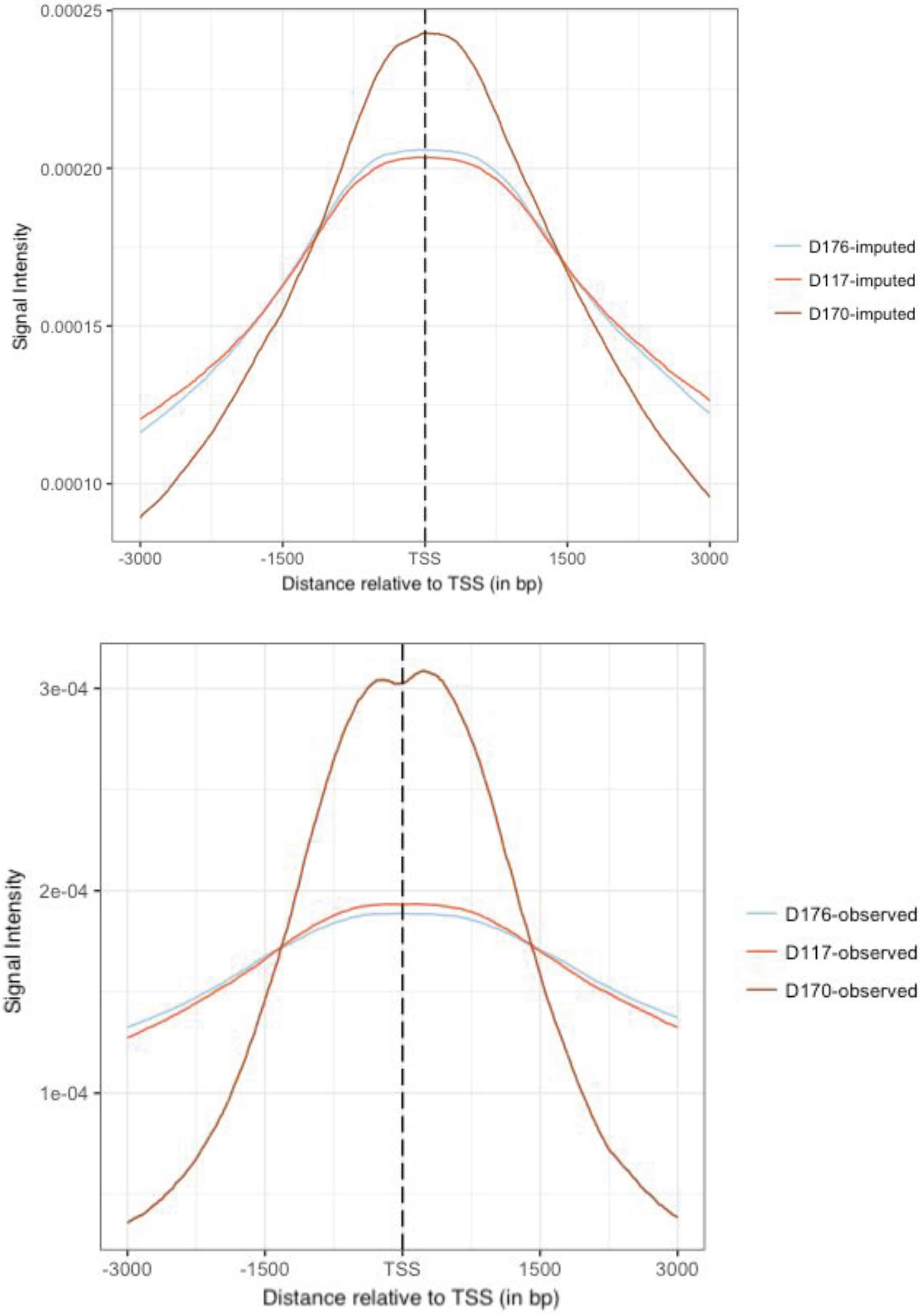


Figure 3.2 Peak overlap with TSS I overlapped peaks to the nearest TSS position within +/- 3kb window. The imputed peaks overlapped closer to the TSS. Results correspond to each respective sample (color code) for the H3K27ac mark. The top panel contains the imputed peaks, while the bottom contains the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

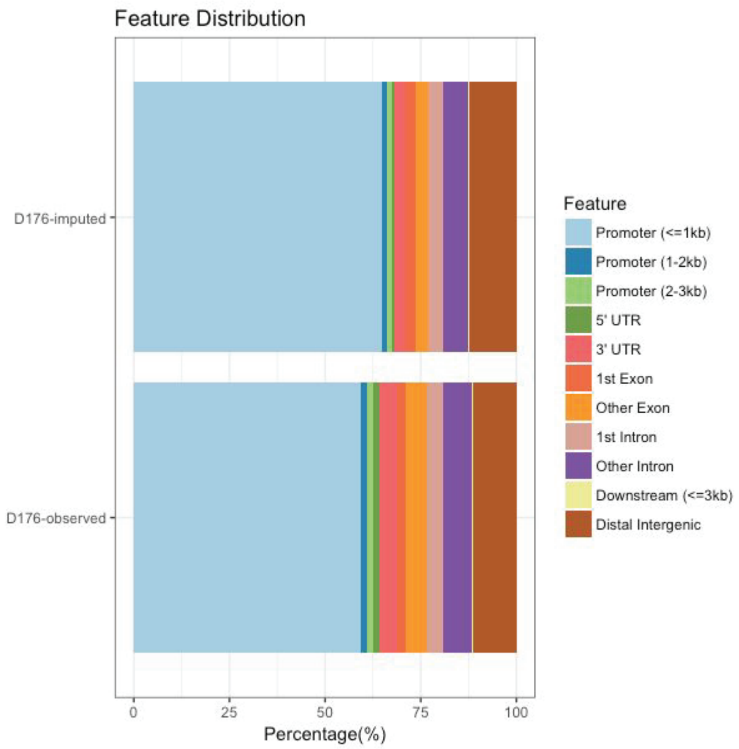
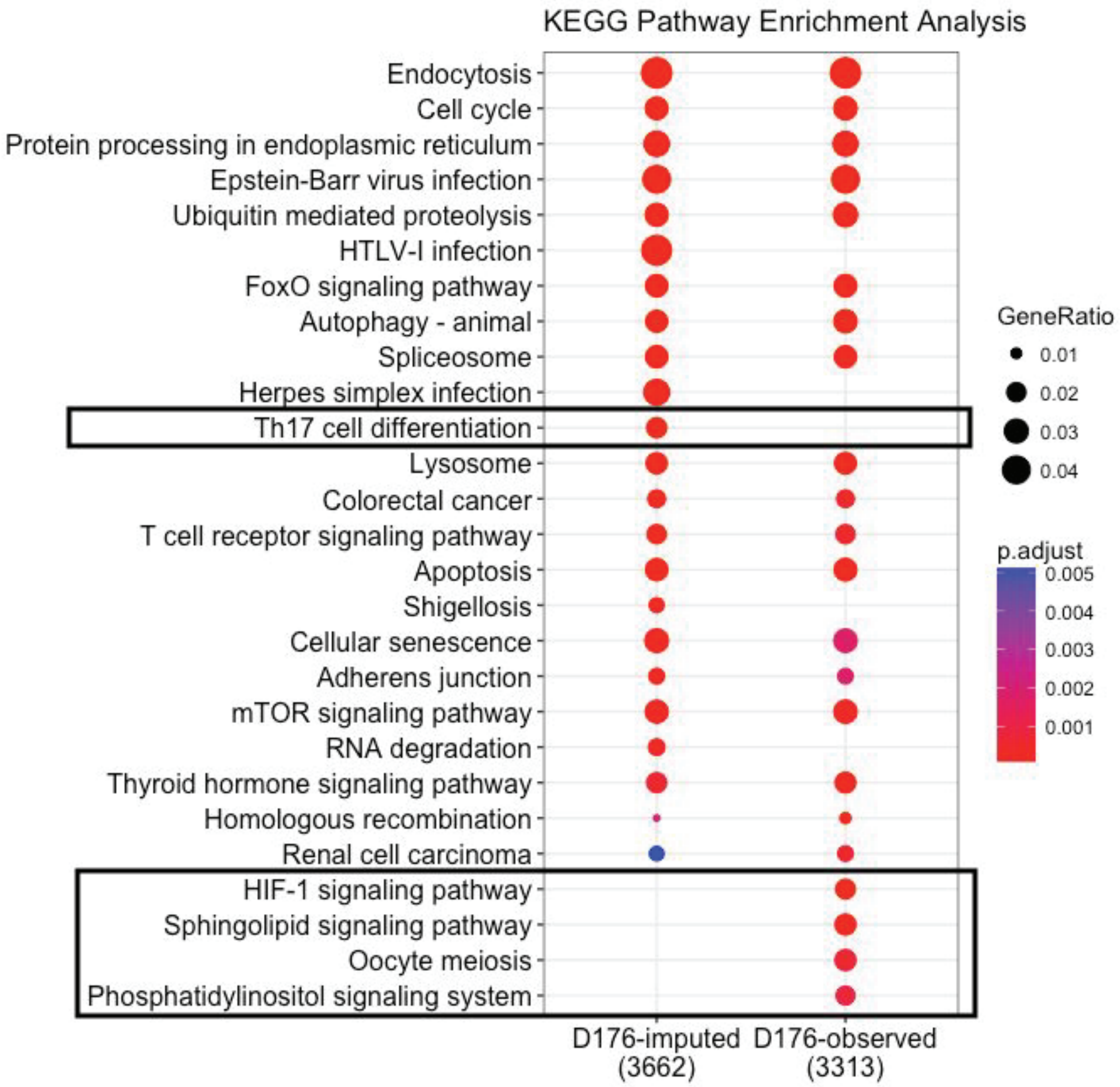


Figure 3.3 Peak annotation relative to gene features. Peaks were annotated by distance to a TSS using UCSC HG 38 known genes (Hsu et al. 2006). When there were multiple peak annotations, the annotation within closest distance to the TSS was chosen. This distribution was able to capture more promotor 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

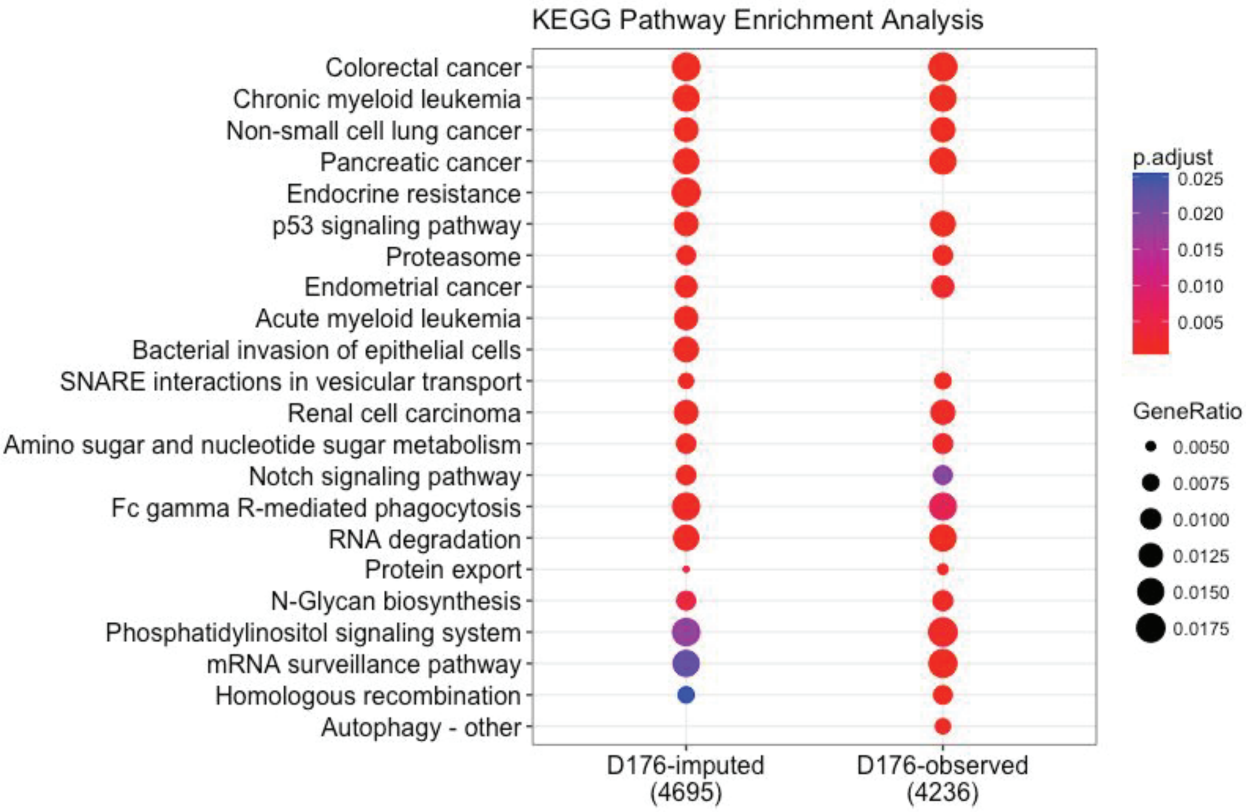


Figure 3.4 KEGG Pathway Enrichment analysis Genes within 3kb to ChIP-seq peaks were annotated via pathway enrichment analysis using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Circle areas represent the gene ratio, which is proportional to the number of genes that overlap a specific 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.

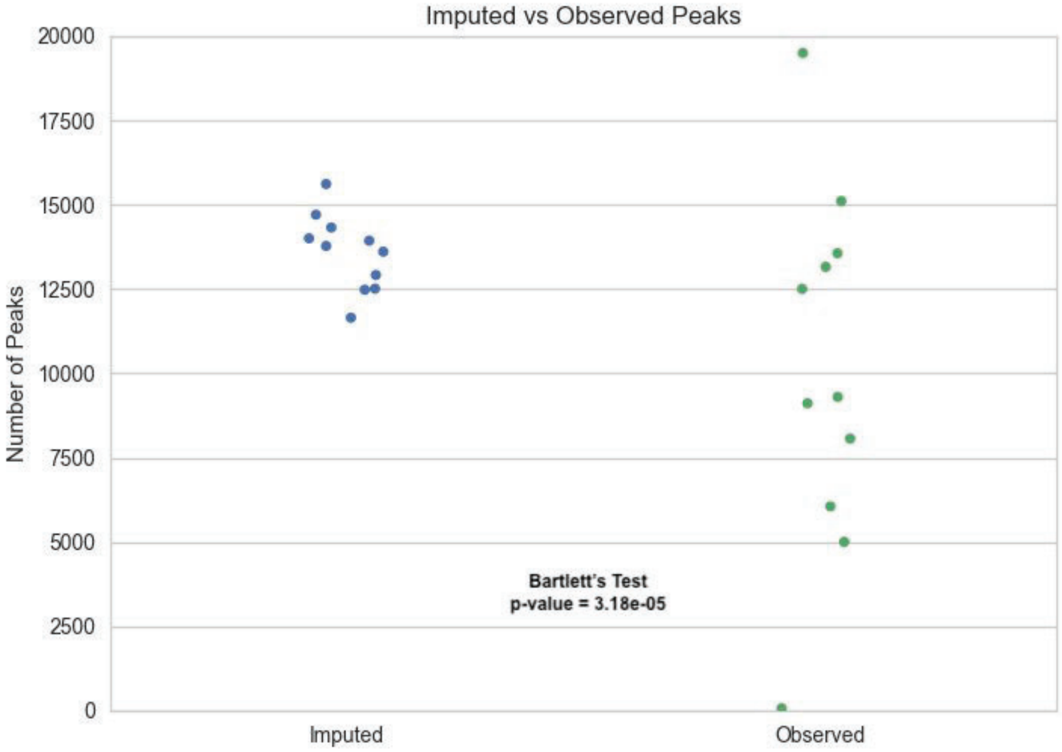
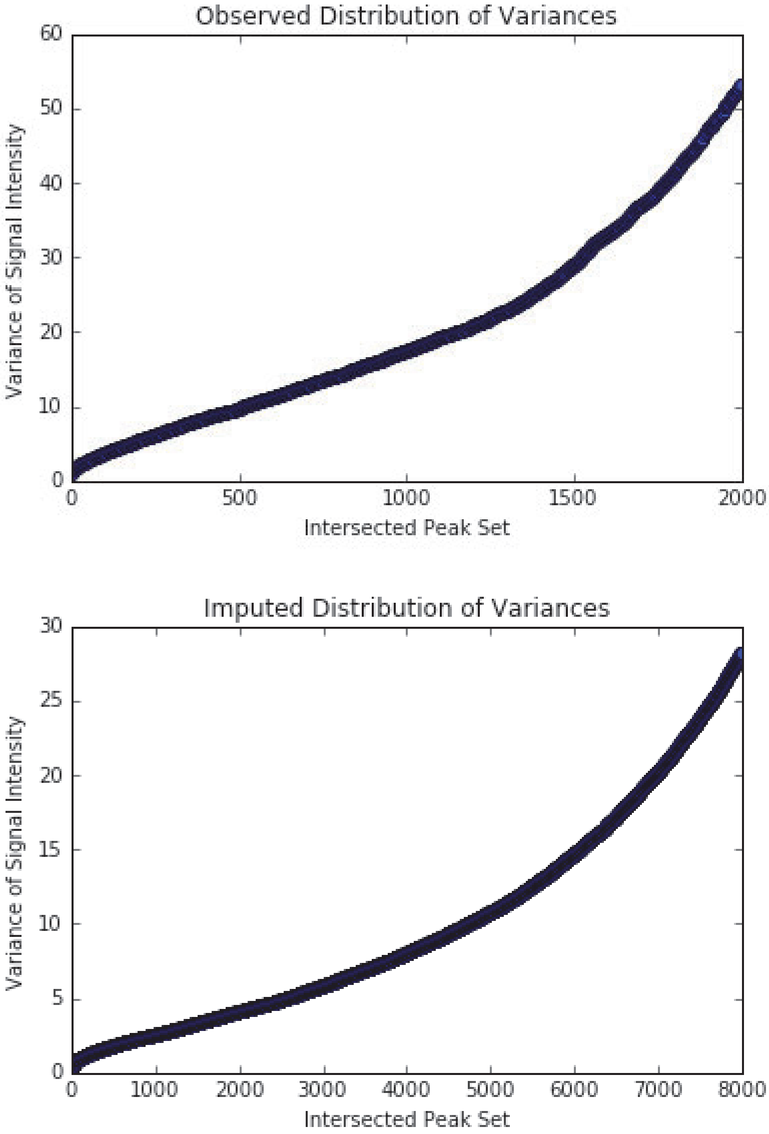


Figure 3.5 Comparison of variance in imputed and observed data. The total number of peaks per sample called across 11 ChIPmentation samples for the observed and imputed data. Bartlett's test for equal variance revealed a p-value of 3.18 x 10^{-5}.

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
0
10
20
30
40
50
60
0
500
1000
1500
2000
Intersected Peak Set
Imputed Distribution of Variances
Variance of Signal Intensity
0
5
10
15
20
25
30
0
1000
2000
3000
4000
5000
6000
7000
8000
Intersected Peak Set

B)

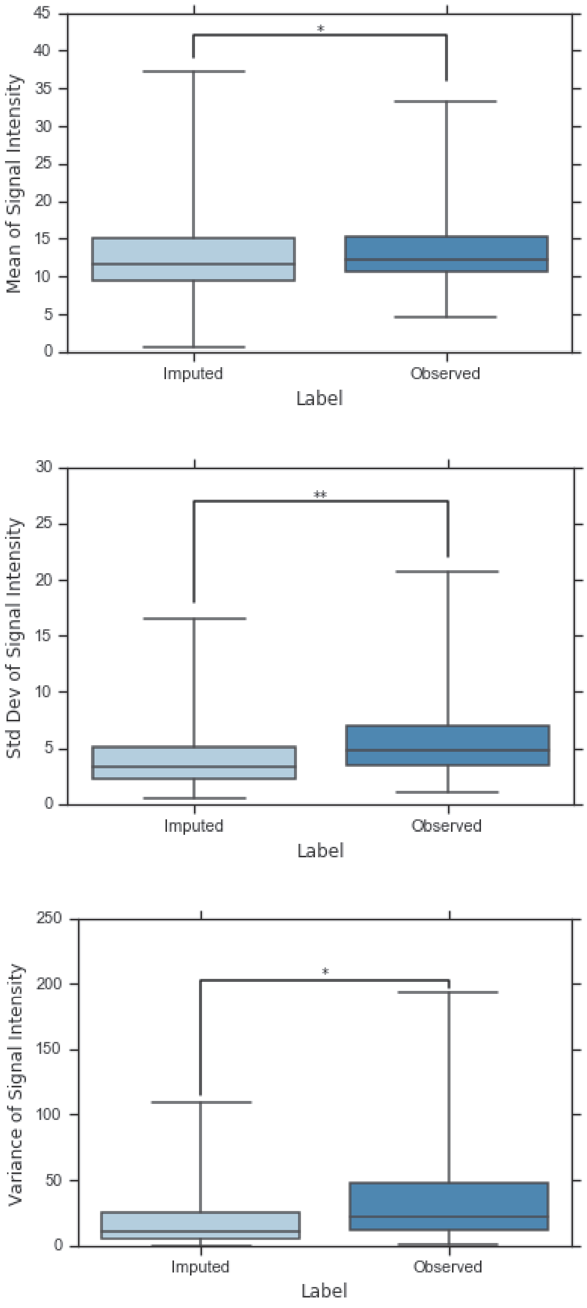
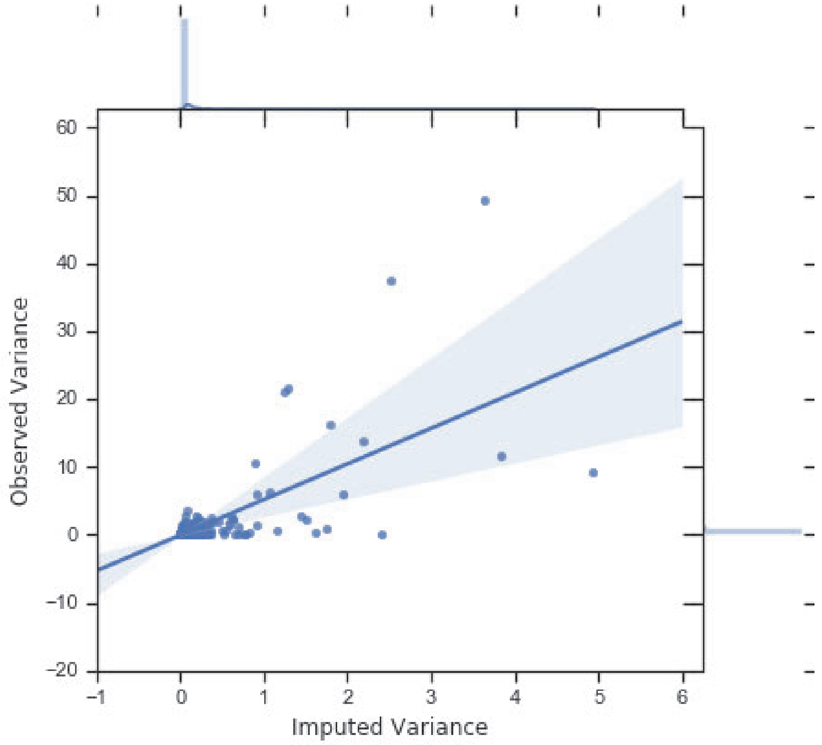
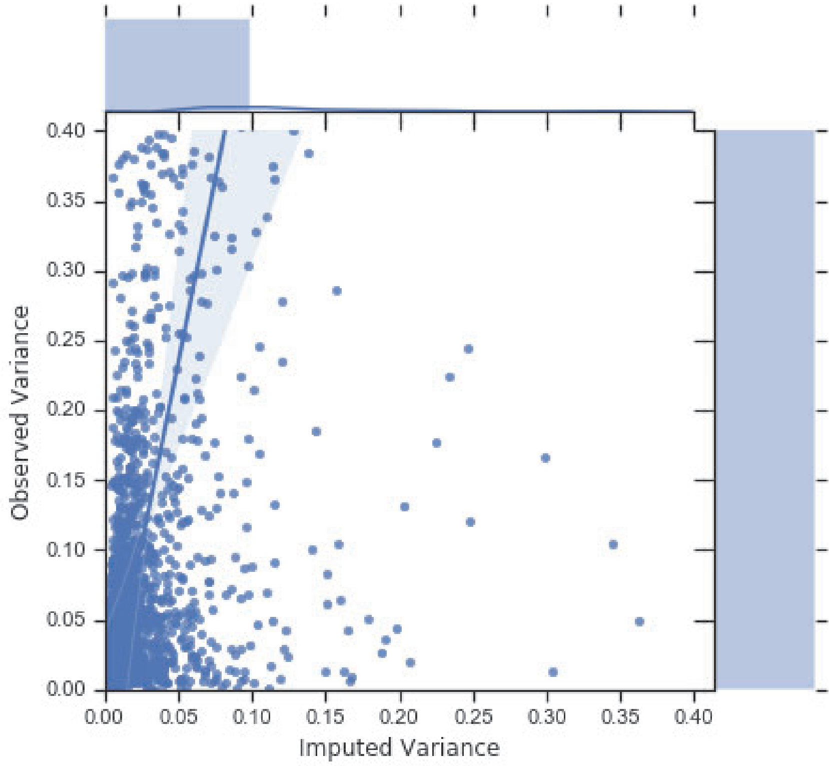


Figure 3.6 Distribution of variance for imputed and observed peaks. I identified peaks that were shared between observed and imputed data by intersecting peaks from 11 samples (minimum overlap of 20 percent). **A)** The distribution of peak variance for signal intensity for 80 percent of the peaks. **B)** 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)


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

B)


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

C)

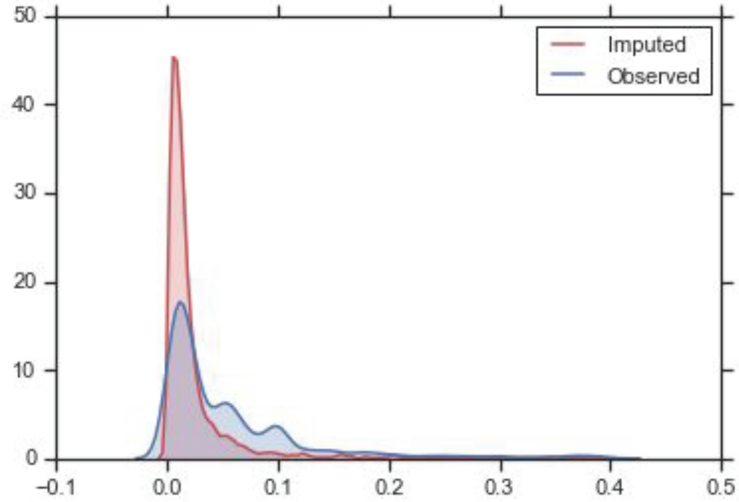
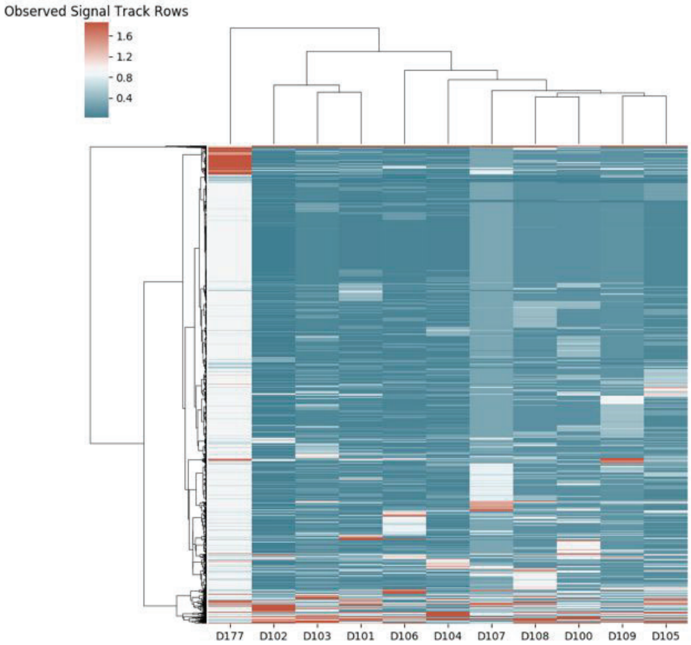


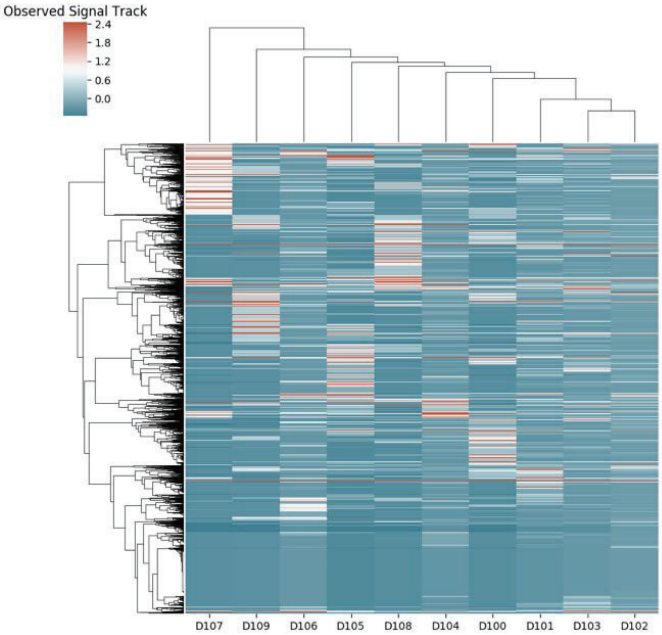
Figure 3.7 Variance between imputed and observed signal track I decomposed observed and imputed signal intensity tracks into 100,000 base pair bins and calculated the variance in observed and imputed signal intensity across the 11 biological replicates. Variance was calculated using the same bin at the 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.

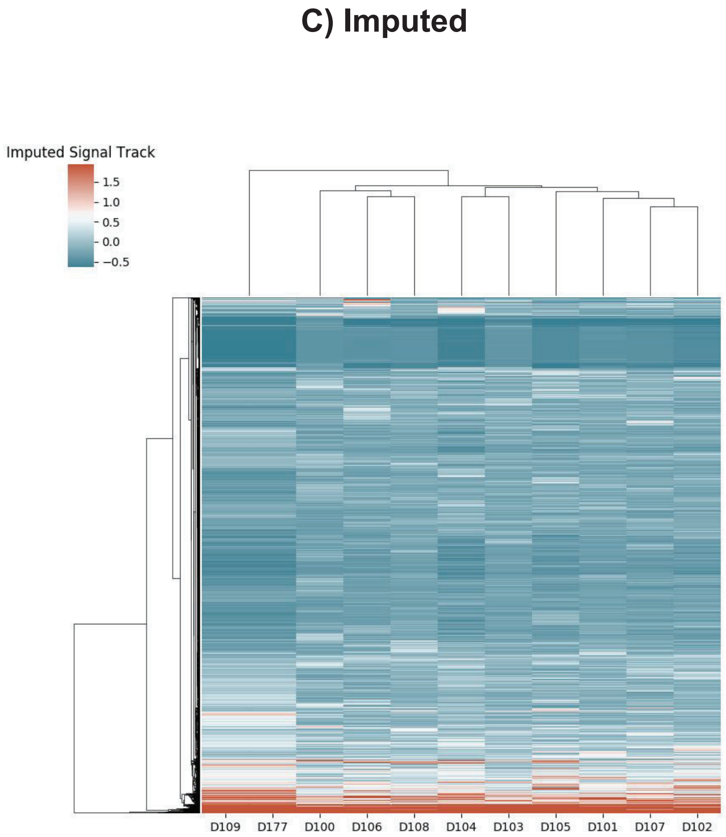
I wanted to understand what was driving the differences between observed and imputed signal tracks. In order to answer this question, I decomposed the observed and imputed signal track for each sample into 100,000 base pair bins. The imputed signal tracks varied far less than the observed tracks **(Figure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

A) Observed



B) Observed without D177





C) Imputed

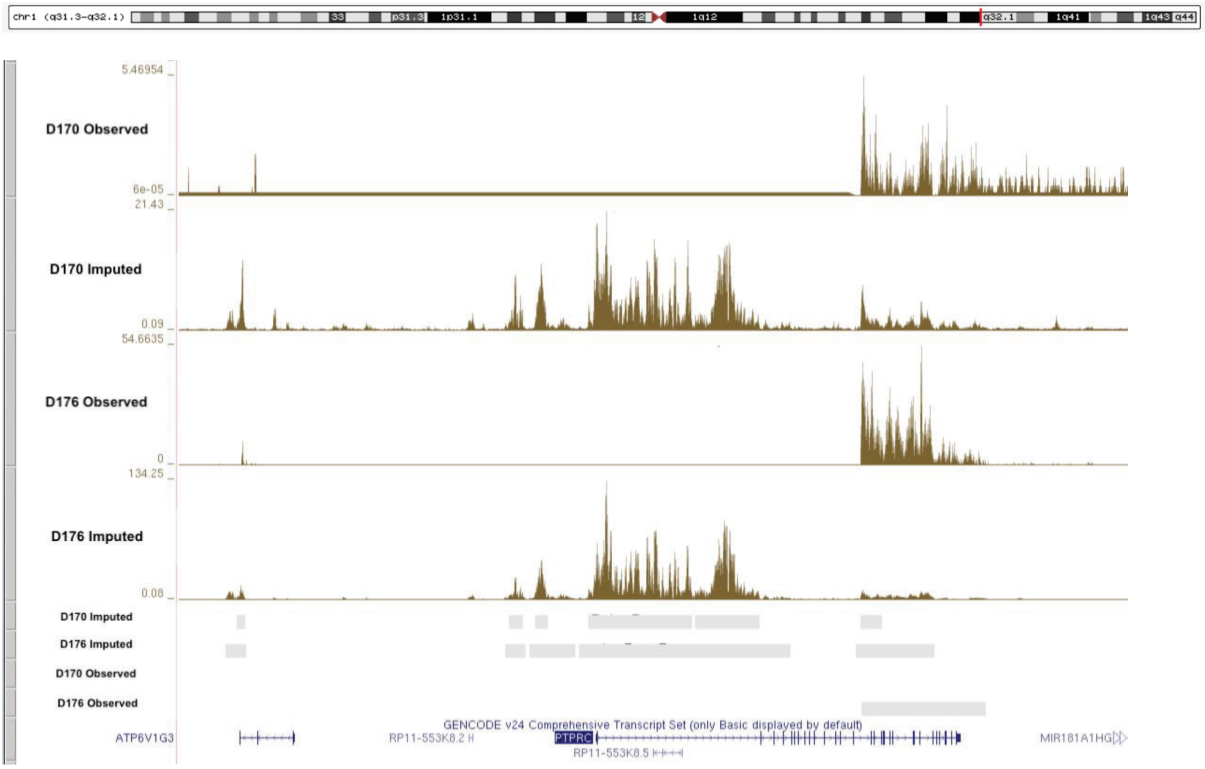
Figure 3.8 Heatmaps of Unweighted Pair Group Method with Arithmetic Mean clustered imputed and observed data. To visualize the similarity between the observed and the imputed data I performed Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clustering on signal intensity in 100,000-bp bins throughout 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

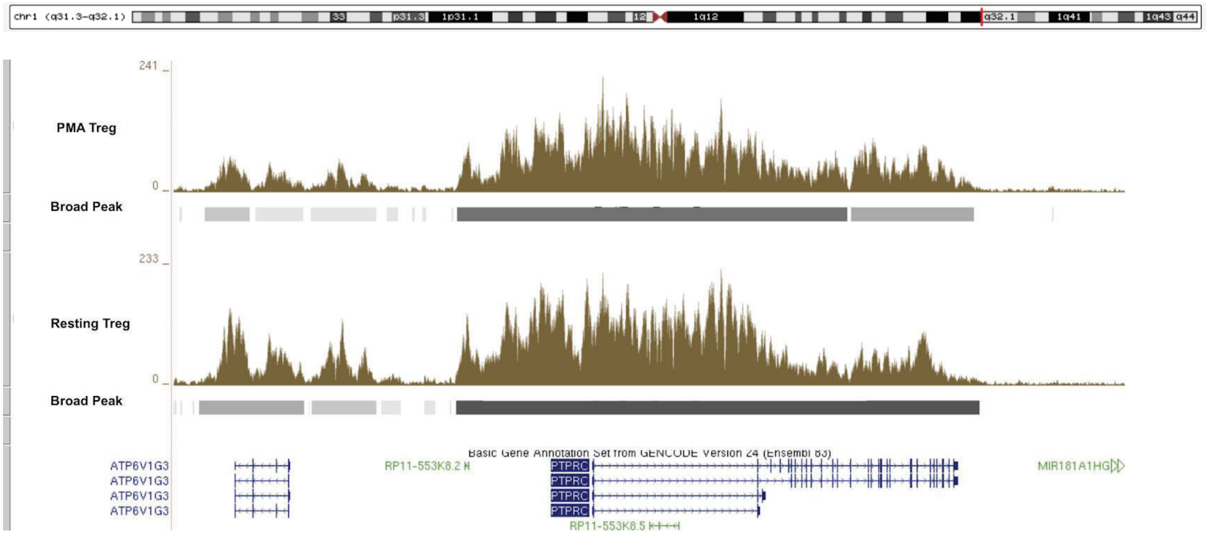


Figure 3.9 Recovery of signal through imputation. Samples D170 and D176 are two individuals for which I observed a loss of intensity signal over a Treg critical CD45 gene, PTPRC for histone mark H3K27ac. The location of this gene is at chromosome 1:198,639,040-198,757,283. **A)** Upon imputation the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

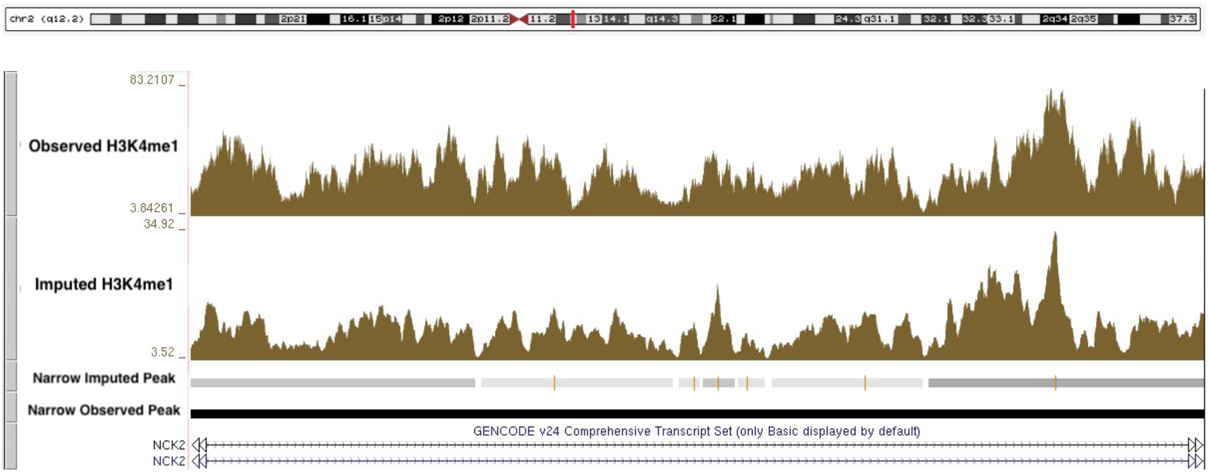


Figure 3.10 Imputation disarranged narrow peaks The sample has been called with narrow peaks, which are the bands below the signal tracks. The imputed peaks have been

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.

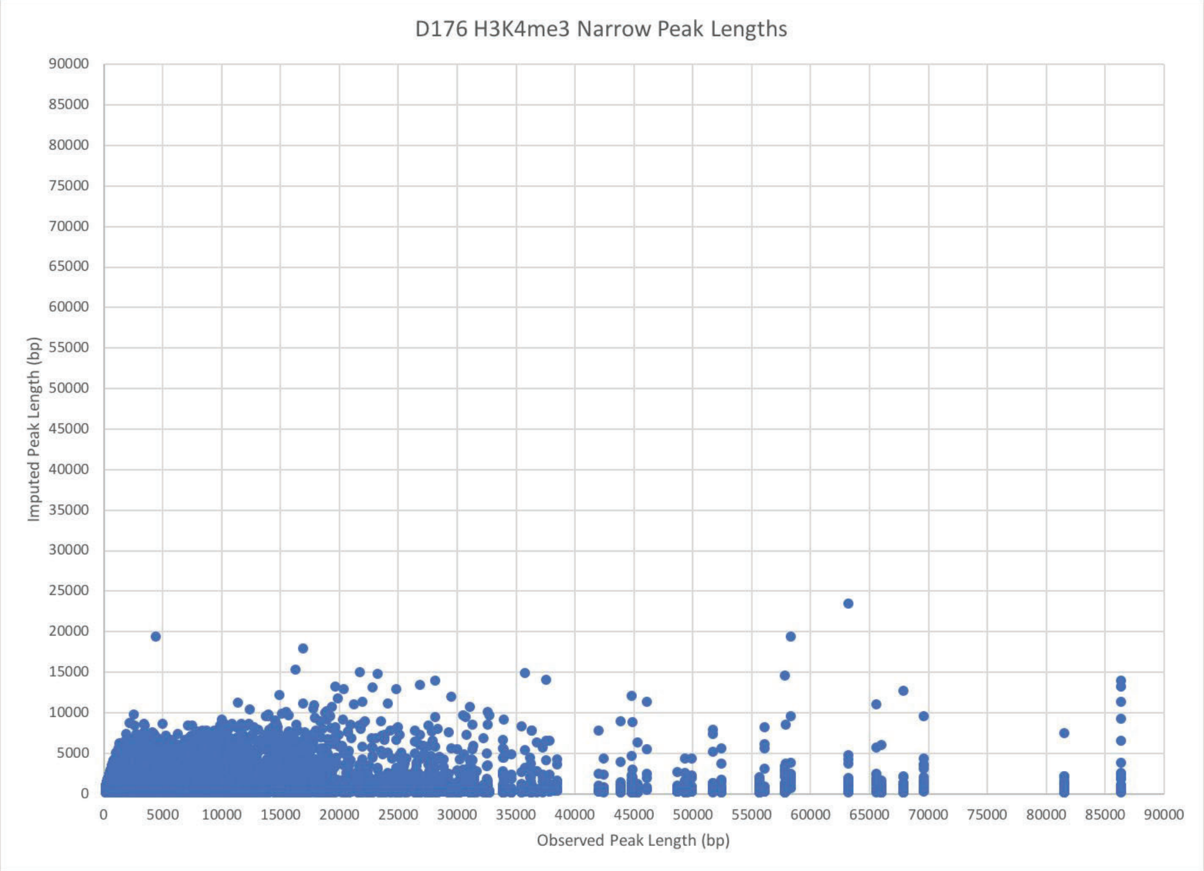


Figure 3.11 Comparing imputed and observed peaks with a narrow peak caller
Narrow peaks were called for both H3K4me3 histone marks for the same sample (D176). After calling, peaks were intersected if the overlap between the peaks was larger than 20 percent. Each of the peaks for the mark would have observed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

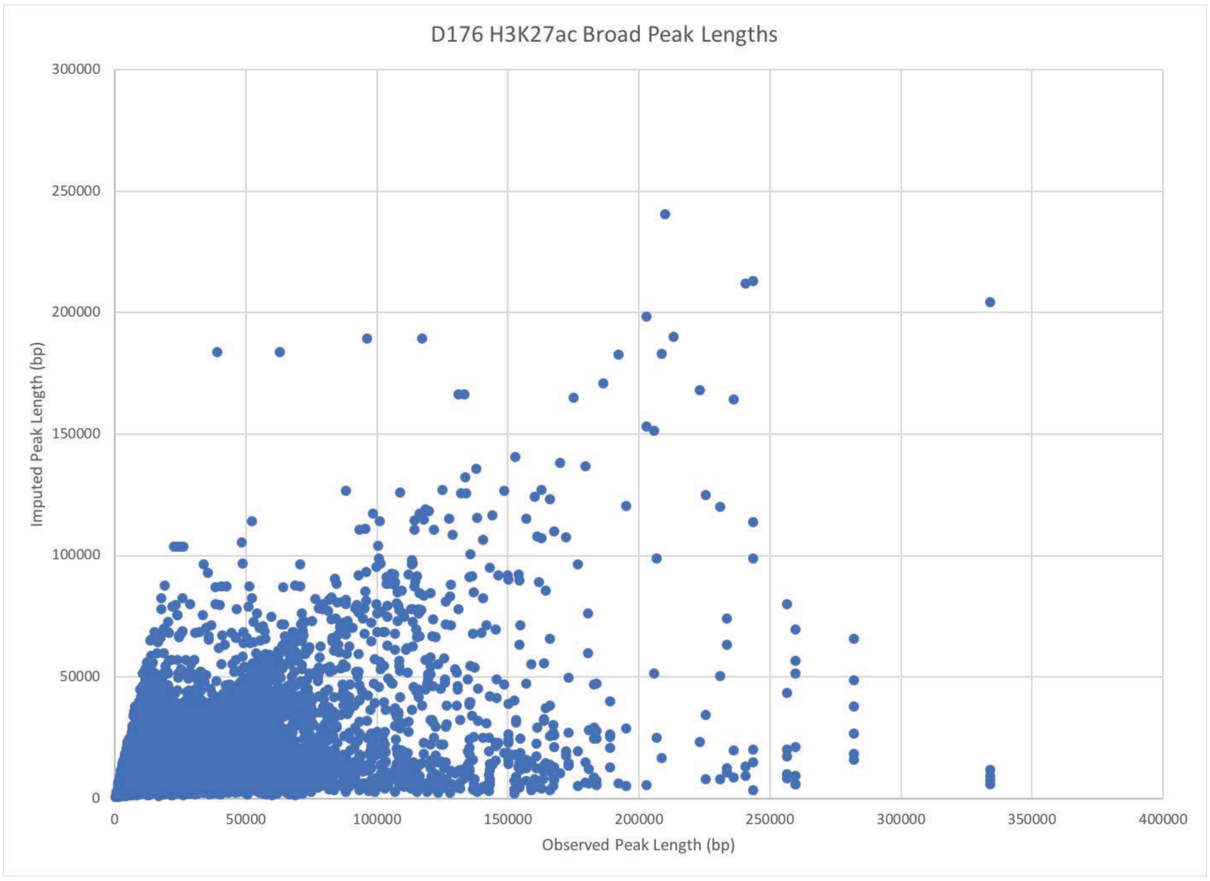
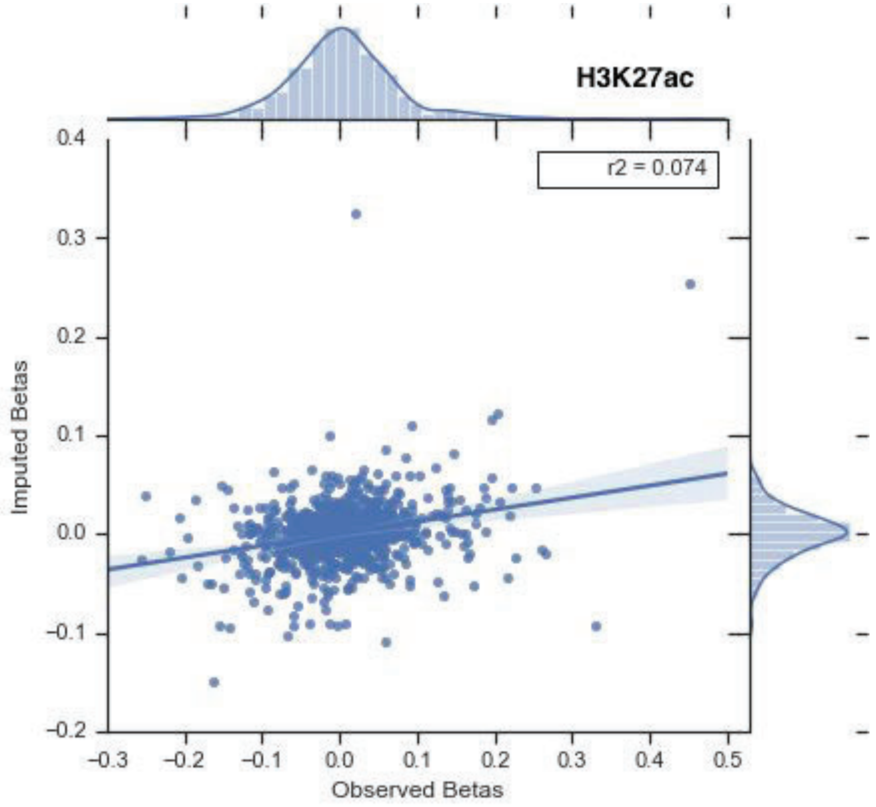


Figure 3.12 Comparing imputed and observed peaks with a broad peak caller Broad peaks were called for both H3K27ac and histone marks for the same sample (D176). After calling the broad peaks, the imputed peaks would either increase in length or decrease in length. Namely, the shattering seems to disappear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

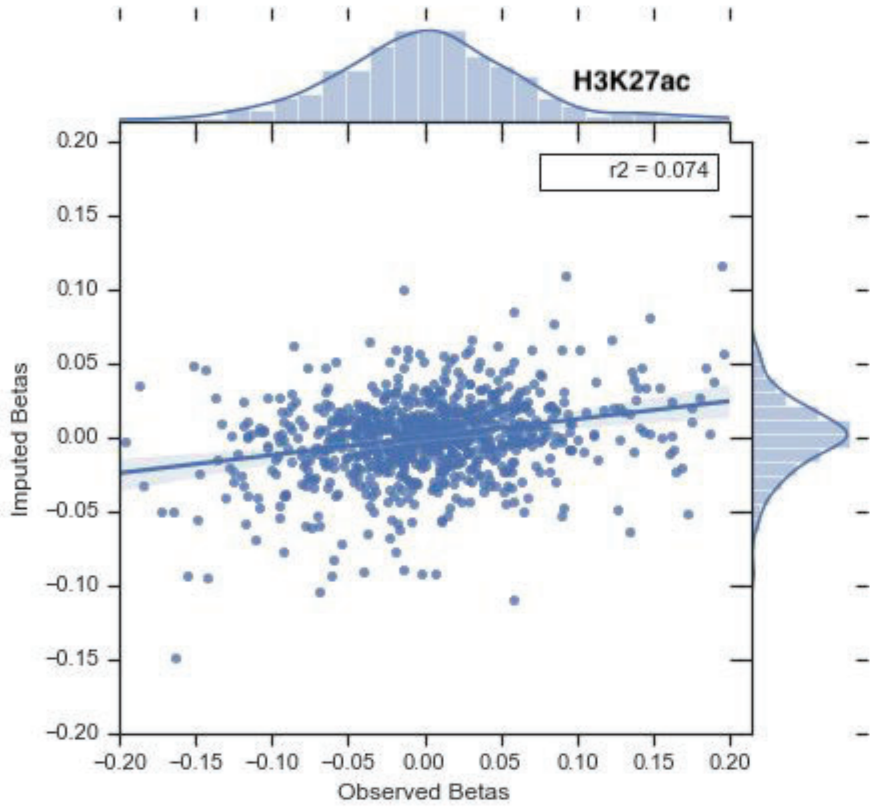
I then performed the same analysis using a broad peak caller, and intersected the observed and imputed peaks. Peaks obtained in this way seemed to gain length in the imputed data as well as have similar characteristics to the observed peaks **(Figure 3.12)**. It is therefore recommended to use a broad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

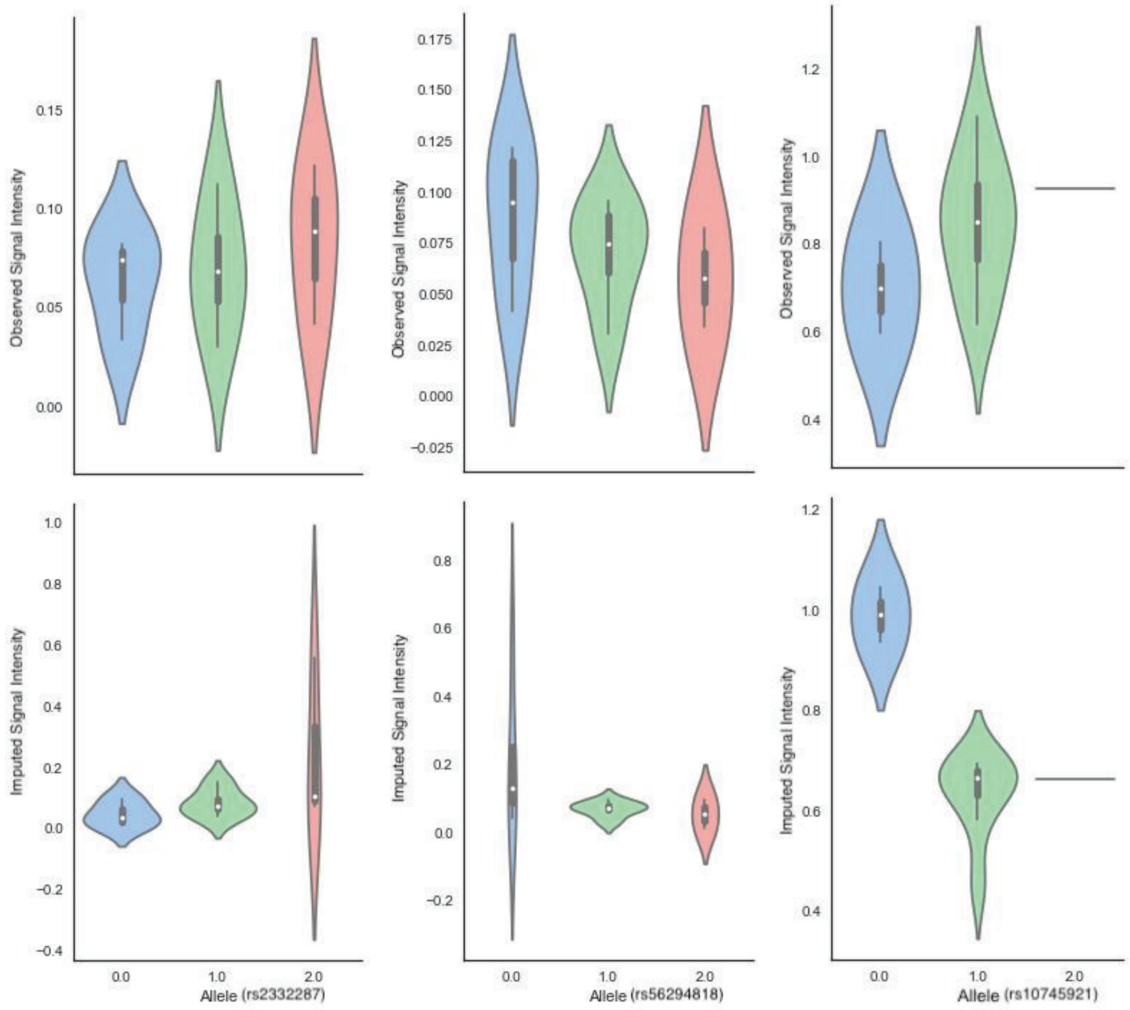


Figure 3.14 eQTL distributions of beta values between observed and imputed signal intensity values Distribution of the signal intensity values from 11 individuals for selected examples of QTL effects after the imputation. In the first example the outlier is not very strong, however the latter examples 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.

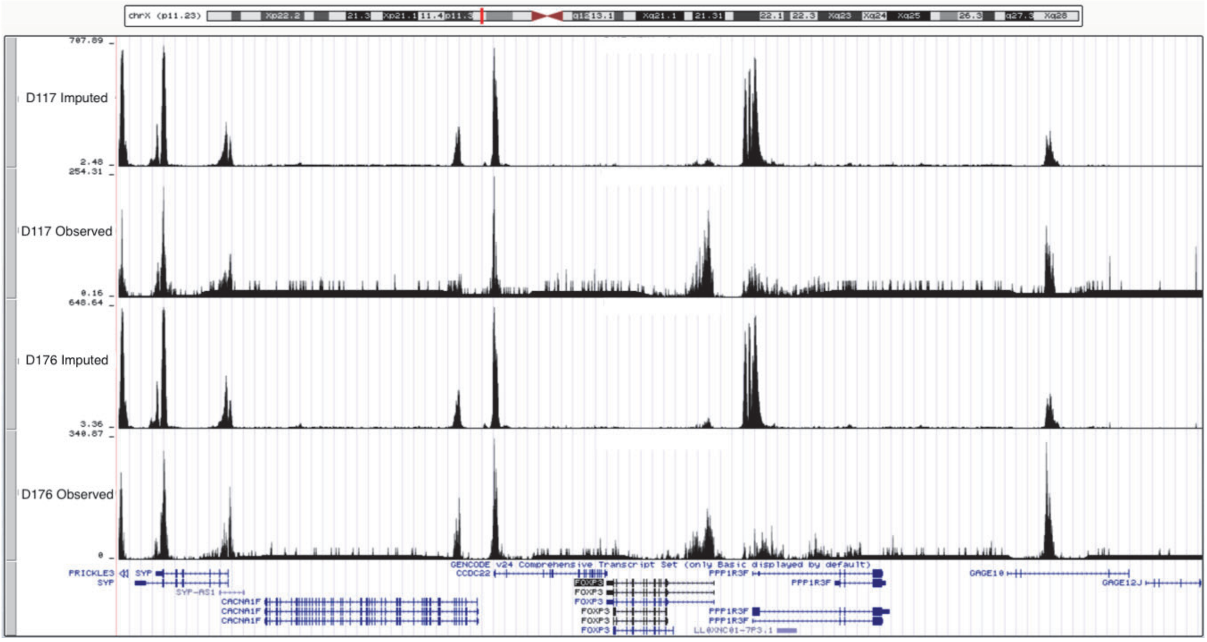
I therefore concluded that ChromImpute does not preserve the genotypic variability, as the genetic effects are often dampened or non-existent.

3.7 Imputation applied to understand Treg biology

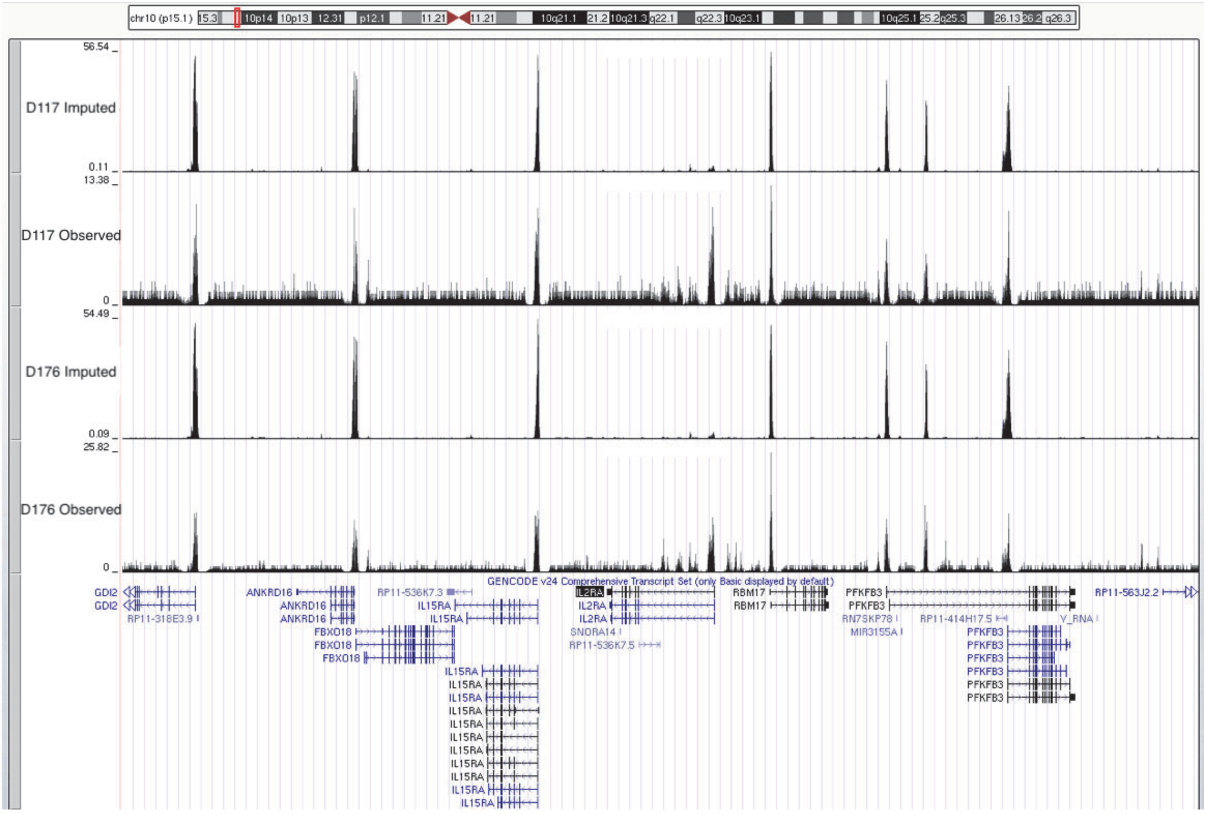
I was interested to see if imputation was able to recapitulate signal specific to Tregs. The challenge of capturing specific signal within a rare cell type is acutely felt, and I wanted to see if imputation amplified overlap for specific Treg genes, as well as any new genes.

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.

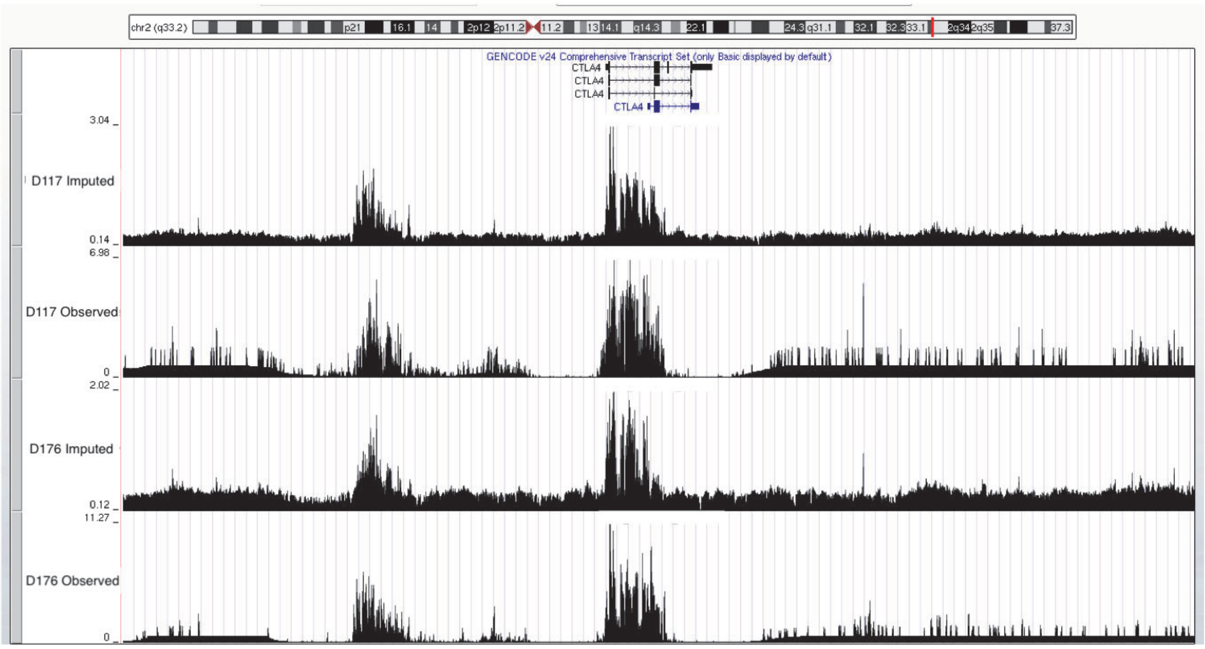
A)



B)


chr10 (p15.1)
D117 Imputed
56.54
0.11
D117 Observed
13.38
0
D176 Imputed
54.49
0.09
D176 Observed
25.82
0
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
GDI2
RP11-318E3.9
ANKRD16
FBXO18
RP11-536K7.3
IL15RA
IL2RA
SNORA14
RP11-536K7.5
RBM17
PFKFB3
RN7SKP78
MIR3155A
RP11-414H17.5
Y_RNA
RP11-563J2.2

C)


chr2 (q33.2)
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
CTLA4
D117 Imputed
3.04
0.14
D117 Observed
6.98
0
D176 Imputed
2.02
0.12
D176 Observed
11.27
0

D)

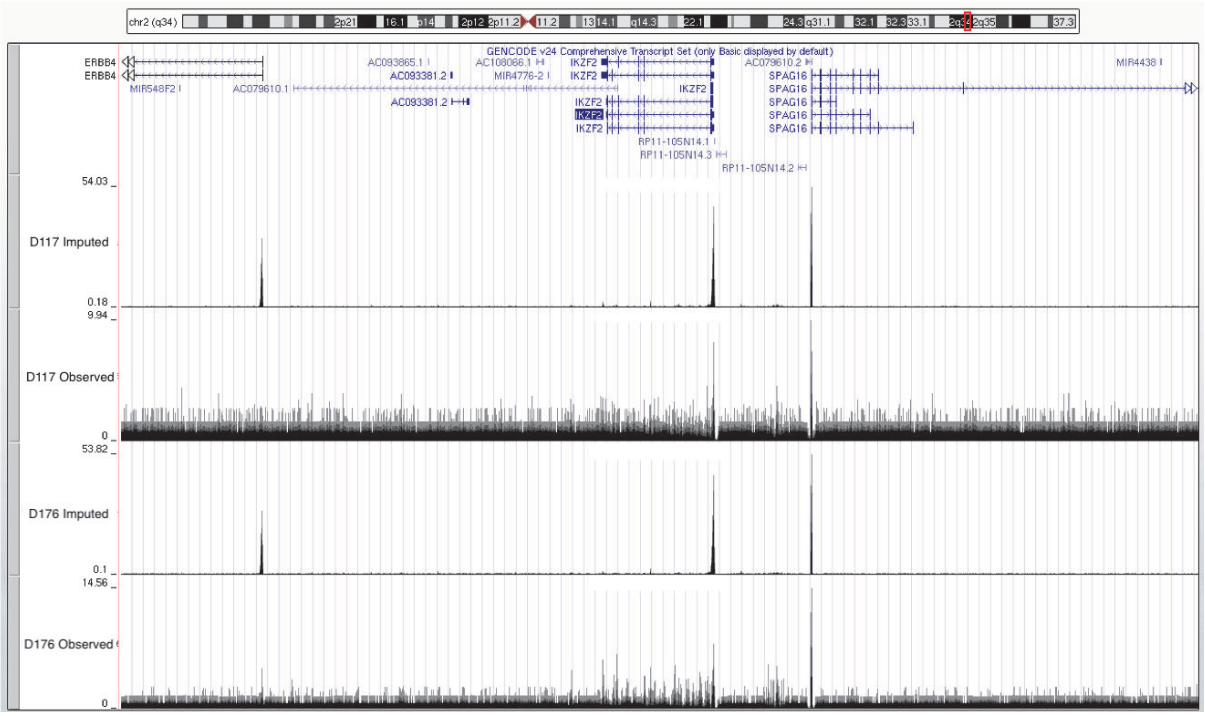


Figure 3.15 Comparison between observed and imputed signal track for key TReg genes Investigation of the key Treg genes are maintained or boosted by the imputation. A selection of 2 different Treg samples for the histone mark H3K4me3. Plots here compare the signal track of for the two samples with observed 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