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.