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