Functional annotation of Rare and Low Frequency variants of Swedish population on putative enhancer regions identified by HiCAP methods

Abstract

GWAS have identified most of variants associated to complex disease to non coding variants. The functional impact of these variants are poorly understood. The non-coding variants are studied using Hi-C and its variant methods such as HiCAP which brings genome into close proximity leading to regulating of gene expression. In current study, we have identified common, low frequency and rare variants one of HiCAP data on two replicates of Bicuspid aortic valve (BAV) heart disease. The current analysis shows that Rare variants are enriched in promoter-mediated putative enhancer. Upon further annotation of these regions with non coding functional markers of DNA hypersensitivity, Histone modification and transcription factor binding sites from public data set shows 23% these enhancers are enriched with at least one of these markers. Furthermore, GO term analysis on enhancers with all of these markers points to one of regulation of endothelial tube morphogenesis genes as one with high enhancer changes.

1 Introduction

Currently genome wide association studies are targeted to common variants that are present in higher frequency in a population, which had lead into hypothesis of common variant common disease hypothesis. However most of common variants identified so far are within the non-coding regions that are difficult to link with disease. On the other end of spectrum, rare variants: less common in population, are less studied due to large sample requirement and higher cost of sequencing. However, recent studies have shown these genetic variants have higher effect on common diseases as well.

One of such method to study these variant using HiCap (variant of Hi-C) methods, that identifies promoter-anchored interaction between variants that are thousand of bases apart from each other. Thus providing connectivity information from GWAS variants with potential genes. HiCAP method generates a genome-wide maps of promoter-anchored chromatin interactions with close to single-enhancer resolution. (cite HiCap methods). Enhancers are the cis acting regulatory elements of non coding genome which

are essential for expression genes. Thus, it would be variation within enhancer region could potential have impact on gene regulation thereby influencing complex diseases.

BAV are the most common type of aortic anomaly that are major common cause of heart disease in adults. BAV are heritable traits with high influencing males than females , however genetics is poorly understood with no clear one gene influencing disease, therby speculating effect of different environmental, genetic and epigenetic factor playing the role in disease.

In the current study we have imputed genetic variants from recently published 1000 Swedish population in promoter mediated putative enhancer from HiCAP. The genetic profile of rare and low frequency variants in interaction data from are from replicates from BAV heart disease patient. We observed overlap of these enhancer with genetic variants to different functional elements from Chipseq atlas for HUVEC cell line which are primary endothelial cell-lines.

2 Materials and methods

2.1 Data acquistion

Whole genome sequencing variant file (vcf) was accessed from swedgen frequency (cite...) data (https://swegen-exac.nbis.se downloads) of v2. As reported these dataset includes the highest quality genetic map of Swedish population From the resulting vcf file, the Snp data set was filtered using a vcftools (cite..) in order to separate snp and indel. Additionally, the variants tagged as "PASS" from GATK (cite..) was further processed . Further the dark listed genome region (Heng li) was removed from subsequent analysis.

Additionally , for functional annotation of enhancer regions chipseq data for DNAise hypersensitivity experiments (DNAise HS), H3Kme1 and H327Ac , and transcription factor experiments for HUVEC dataset was accessed. The threshold for experiments was set up at 100 and all the peak files were downloaded. The DNAise sites have lost their condensed chromatin and are exposed for expression. H3Kme1 and H327Ac are the active markers of putative enhancers and transcription factor are protein complexes that bind to genome for gene expression. These non coding DNA elements are putatively functionality markers of potential enhancers.

2.2 Definition of Rare and low frequency variant in the population

The variants from the Swedish population was classified into three separate categories i.e Rare, Low frequncy and common based on the allele frequency in the population The variant classification were on the frequency such that variant with MAF>0.05 were classified as "Common", Low frequency with $0.05 < \text{MAF} < 0.01$, Rare variants < 0.01. However in the rare frequency variants were have removed that private variants that were present within one individual either in homozygous or heterozygous condition

2.3 HI-CAP interaction dataset

HiCAP experiments on two replicates from BAV patients were performed. The data were preprocessed with each PromoterEnhancer (PE) interaction having at greater than three supporting pairs of p-value < 0.001 as set in earlier experiments. The final output of the PE interaction data consists of information of Promoter in genome, corresponding genes, enhancer position in genome and supporting pair information for each interaction in two replicates.

Following the interaction dataset, I developed customized python script $VCFmanipulation.py$ which takes the interaction file and annotates and subsets the enhancer and promoter regions into the rare, low frequency and common regions. Furthermore, I also wrote customized python script $genome_tf.py$ further annotates the DNA elements in these regions.

2.4 GO term enrichment status of P_E interaction mediated genes

GO term of biological processes (BP) that were curated from the GO database were down from QuickGo (cite..) . Additionally while downloading the BP date we considered only the terms that were fulfilled the criteria of manual experiment evidence given in the QuickGO database. We used this criteria in order to limit or gene ontologies analysis to relatively functional genes which have the experimental validation. A customized python $Gene\_ontology.py$ was made in for the following analysis as well.

3 Results

3.1 Mapped read in each platform

3.2 Average number of called variants.

3.3 Average coverage of called variants

3.4 Average genotype quality of called variants

3.5 Discordant variants

4 Discussion