Functional Annotation of Rare and Low Frequency variants on putative enhancer regions identified by HiCap methods

Sailendra Pradhananga
Science for Life Laboratory/School of Biotechnology
Royal KTH institute of Technology (KTH), Swden

Abstract

Enhancers are non coding DNA elements essential for regulation of genes [1]. However they are functional after genome looping which brings distant regions in closer proximity to promoter of genes. Variants in these regions are already implicated in complex diseases. In current study, we have identified common, low frequency and rare variants on HiCap data on two replicates of Bicuspid aortic valve (BAV) heart disease in promoter enhancer interaction genome. The study shows that rare variants are enriched in promoter-mediated putative enhancer. Further annotation of these regions with non coding functional elements with HUVEC cell-line data set reveals at least 23% these enhancers are overlapped with at least one of noncoding markers. Gene Ontology term analysis with promoter enhancer interaction (PEI) genes reveals regulation of endothelial tube morphogenesis as one of the top GO term enriched with Low and rare frequency variants.

1 Introduction

Chip-based genome wide association studies on complex diseased targeted common variants (frequency > 0.05) which had lead to hypothesis of common variant common disease hypothesis [2] However most of common variants identified within the non-coding are unable to explain causality to complex disease. On the other end of frequency spectrum, rare and Low frequency variants, are less studied due to large sample requirement and higher cost of sequencing. Recent studies have shown these genetic variants have higher effect on common diseases as well [2–4]. Furthermore, these non coding rare and common variants have shown to disrupt gene regulation mechanism in promoter-enhancer regions. [5, 6] .

One of method to study of studying looping genome is HiCap (variant of Hi-C) [7]. The method identifies promoter-anchored interaction between genomic regions that spatially, however are thousand kilo bases apart. Additionally, these interaction provides connectivity information of for example enhancers with potential genes.. In relation to other competing technologies, HiCap method generates a genome-wide maps of promoter-anchored chromatin interactions with close to single-enhancer resolution. Enhancers are the cis acting regulatory elements of non coding genome which are essential for regulation of genes transcription. Since enhancer are functionally important region of non coding genome, variations in these regions are of considerable importance in complex diseases [8, 9]

Biscupid aortic valve (BAV) disease is most common type of aortic anomaly in adults [10]. BAV are heritable traits with high influencing males than females, however genetics is poorly

understood with no clear one gene influencing disease, thereby speculating effect of different environmental, genetic and epigenetic factor playing the role in disease [10, 11]. There has been considerable interest to find genetic markers in understanding the prognosis of disease.

In the current study, we have imputed genetic variants from recently published 1000 Swedish population [12] in PEI genomic region from HiCap data of BAV patients. The genetic profile of rare and low frequency variants in PEI are from two technical replicates from BAV heart disease patient. We observed overlap of these enhancer with genetic variants to different functional elements from Chipseq atlas for primary endothelial HUVEC cell lines.

2 Materials and methods

2.1 Data acquisition

Whole genome sequencing variant file (vcf) was accessed from SweGen [12] database website [13]. 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 [14] in order to separate genetic variant into single nucleotide polymorphism (SNP) and insertion/deletion (INDEL) type. Additionally, only the variants tagged as "PASS" from GATK [15] was further processed . Additionally, dubious genome listed in human genome [16] was removed from subsequent analysis.

Additionally , for functional annotation of enhancer regions chipseq atlas [17] data for DNA hypersensitivity experiments (DNAse 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 DHS 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 frequency and Common based on the allele frequency in Swedish 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 HiCap interaction dataset

HiCap experiments on two replicates from BAV patients were performed. The data were preprocessed with each Promoter−Enhancer Interaction (PEI) pair having at greater than three supporting pairs of p-value < 0.001 in earlier experiments. The final output of the PEI data consists of information of promoter and corresponding genes, putative enhancer position in genome and supporting pair information for each interaction in two replicates. Following the interaction dataset, a customized python script *VCFmanipulation.py* was developed . The script takes the interaction file and vcffile as input and subsets PEI regions into the rare, low frequency and common regions. Furthermore, the output from this file is inputted to python script *genome_tf.py* further annotates the DNA elements in these regions based on annotation files donwnloaded from Chipseq atlas.

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 [18] . 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 Rare variants are enriched in putative Promoter - Enhancer Interaction

There was 35million variants that were tagged as "Pass" all 1000 Swedish genome population. From that, we have identified around 30.2 million SNPs and 4.4 million indels in in the population that are tagged as "Pass" from GATK filter.

The preprocessed promoter-enhancer list contained 33,323 unique enhancers regions in Bicuspid aortic valve. This data contains of promoter regions and corresponding enhancer regions of 2 replicates from BAV cell. We found on average of 20.38 and 13.93 interaction enhancers change in replicate1 and replicate2 respectively.We observed that enhancer had different length distribution as shown in table 1 and fig 1. This length distribution depicts that most of our putative enhancers are within the range of 50-1Mb base pairs which is in par with the Hi-C methods

Enhancer_length	Counts type
< 50kb	82591
$50 - 500$kb	20700
500kb -1Mb	3334
> 1MB	1030

Table 1: Number of enhancer of different length

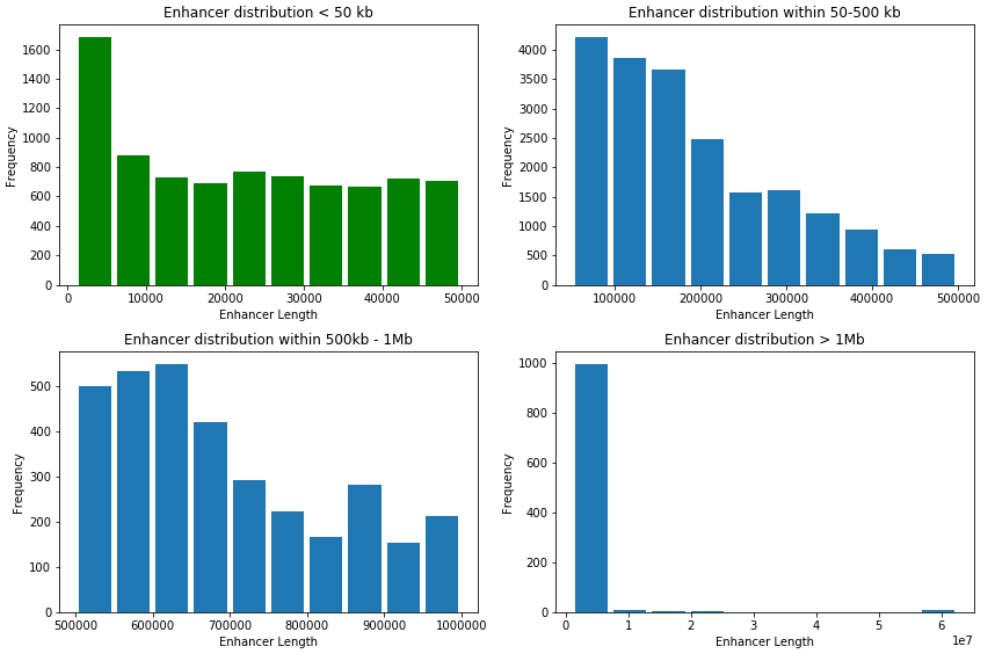

Figure 1: Distribution of Enhancer length on PEI dataset. Most of enhancer was between 50 Kb− 1Mb in length

We identified 22,055, 14403 and 22,144 putative enhancer regions in our interaction dataset with common, low-frequency and rare variants respectively . Furthermore, we identified in total 56,891 common , 24,049 low-frequency and 47,281 rare variants enriched in these enhancer regions as shown in table 2 . Interestingly, our enhancer regions have been enriched with rare variants from Swedish population as compared to other variant class fig 2.

Variant class	Enhancer with variants	Total variants type
Common	22055	56891
Low Frequency	14403	24049
Rare	22144	47281

Table 2: Number of different class of variants in PEI enhancers

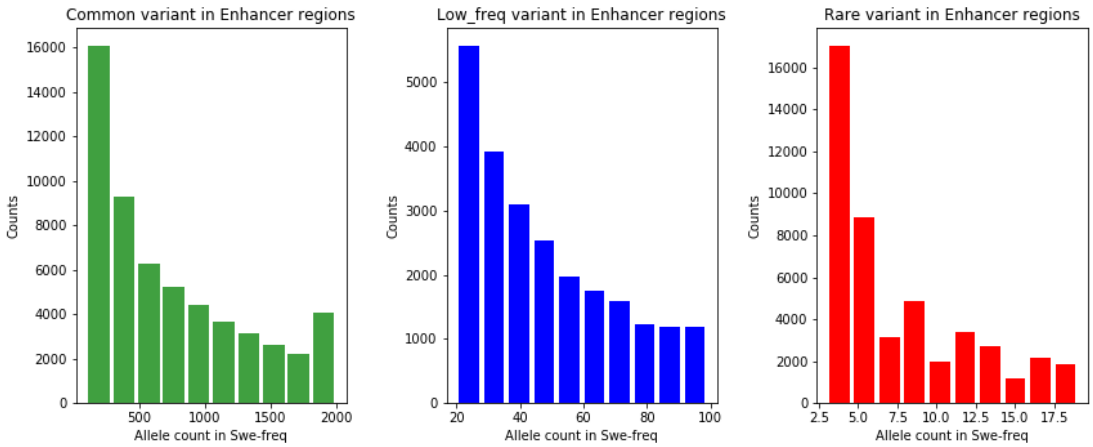

Figure 2: Distribution of different variant class in PEI. The number of allele count in Swefreq Database in x−axis and count in y−axis. Rare variants are enriched with the allele count < 5.

3.2 Status of non coding functional DNA elements of variants embedded enhancer regions

Upon overlapping with the chipseq atlas experimental data with different DNA elements, we all the DNA elements are present in our enhancer list in all classes as shown in table 3.

class	at_least_one	DNS	HM	TFs	DNS+HM	DNS+TFs	HM+TFs	ALL
Common	5087	3350	3029	3416	2042	2487	1744	1565
Low Frequency	3407	2276	2038	2356	1393	1736	1229	1095
Rare	5178	3440	3097	3524	2114	2587	1820	1638

Table 3: Different DNA elements in enhancer of variant classes Common, Low frequency and Rare

Additionally, when we compared types of DNA elements we observed hat we have at least one of functional markers in about (5087/22055) 23% of putative enhancer in all classes of variants as shown in fig 3. However, it has to be considered that we took into consideration one of cell line and these are the markers specific to cell type. Most interesting , we still find at least 1000 putatively, functional enhancers in all variant classes. It would be interesting to observe these functional enhancer regions and dig into Rare and Low frequency variants in these region.Furthermore, these are result only from overlapping with one dataset. This data has to be randomized and overlapped so as to find the putative p-values to our non coding functional elements.

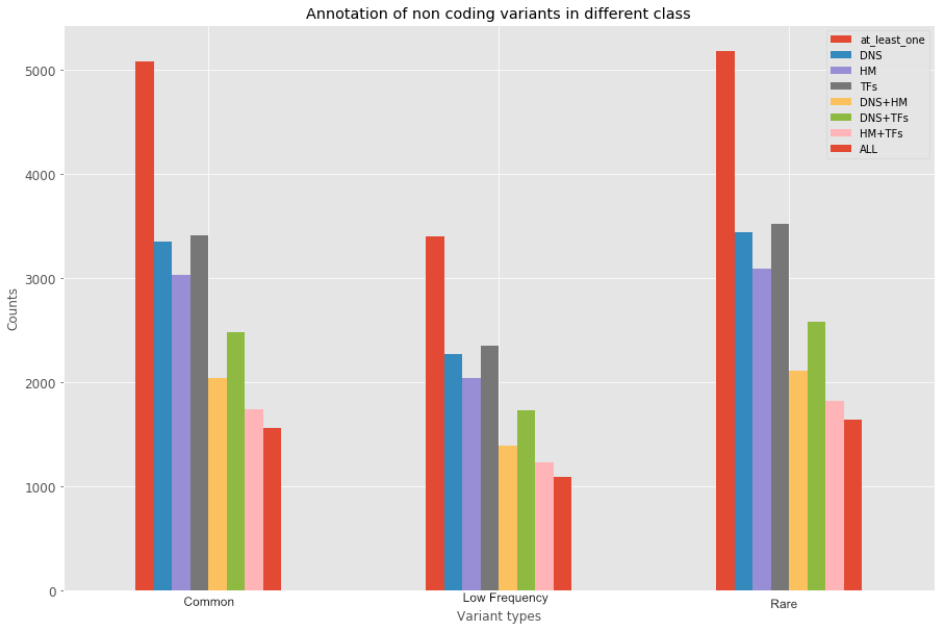

Figure 3: Distribution of different DNA elements in enhancer with variant classes

3.3 GO term enrichment status of PEI mediated genes on variant with all functional markers

We sub selected the enhancers with all the functional marks as we had observed earlier. Upon Gene Ontology analysis of PEI mediated genes on PEI of variant class rare and low frequency on such enhancers resulted some interesting observations. We found 2 enriched GO Terms at PEI more than two. The top GO enriched terms were 1) *GO:1902894 regulation of pri-miRNA transcription from RNA polymerase II promoter* and 2) *GO:1901509 regulation of endothelial tube morphogenesis.*

4 Discussion

In the current report we have presented the status of genetic variants identified from Swedish population in the putative enhancer genomic regions mediated by promoter-enhancer mediated interaction (PEI). These PEIs were called using HiCap experiments in replicates of BAV heart disease patient. Our analysis depicts that rare variants are enriched within our enhancer regions which was as expected since these rare variants also contains the variants which might be presented in one individuals as we haven't corrected for homogenous individuals. However we still find variants which are presented at least greater than five allele counts .

Based on the definition rare, low frequency and common variants, we annotated our enhancer with active enhancers marks. We used publicly available HUVEC dataset for chipseq atlas on DNAse Hypersensitivity. These regions provided the accessibility of genome for further processing . In the same enhancers regions we annotated with active enhancers histone marks (H327ac and H3Kme1) and transcription factor binding sites. The formation of complex between TFs, enhancers and promoter have been already reported as fundamental mechanism for regulation of trancription [19]. At the current state of analysis, we find more than 1000 enhancer regions

which have been all of these functional marks. This is still a lower number and overlapped within one type of cell-line. Based on the this result, we can regards these enhancers are functional . Furthermore, from this primary analysis, we have to calculate enrichment status based on random distance mediated enhancer.

GO-Term analysis have represented regulation of endothelial tube morphogenesis as one of GO with high enhancer activity in rare and low frequency class of variants. Endothelial progenitor cells have been associated as biomarker of cardiovascular disease [20]. There has been elucidation of endothelial cells in relation to inflammation in vascular disorders. Thus, the genes associated i.e *FOXP1* and *FGF1* and their corresponding enhancer could have functional implication in understanding disease prognosis. However these genes and GO term analysis needs further processing on experimental validation.

5 Conclusion

We have used currently available large scale genetic frequency data in homogenous Swedish population to identify the promoter mediated enhancer enriched with rare variants. Upon further annotation of these enhancer with functional markers lead to identification of interesting GO-terms. Although HiCap based enhancer analysis is currently fully mature, we believe that preliminary results points that rare variants could potentially driving the complex phenotypes such as BAV heart disease.

References

[1] D. Shlyueva, G. Stampfel, and A. Stark, "Transcriptional enhancers: from properties to genome-wide predictions," *Nature Reviews Genetics*, vol. 15, no. 4, pp. 272–286, 2014.

[2] E. T. Cirulli and D. B. Goldstein, "Uncovering the roles of rare variants in common disease through whole-genome sequencing." *Nature Reviews Genetics*, vol. 11, no. 6, 2010.

[3] L. Bomba, K. Walter, and N. Soranzo, "The impact of rare and low-frequency genetic variants in common disease," *Genome biology*, vol. 18, no. 1, p. 77, 2017.

[4] A. Saint Pierre and E. Génin, "How important are rare variants in common disease?" *Briefings in functional genomics*, vol. 13, no. 5, pp. 353–361, 2014.

[5] S. Chatterjee, A. Kapoor, J. A. Akiyama, D. R. Auer, D. Lee, S. Gabriel, C. Berrios, L. A. Pennacchio, and A. Chakravarti, "Enhancer variants synergistically drive dysfunction of a gene regulatory network in hirschsprung disease," *Cell*, vol. 167, no. 2, pp. 355–368, 2016.

[6] V. Rusu, E. Hoch, J. M. Mercader, D. E. Tenen, M. Gymrek, C. R. Hartigan, M. DeRan, M. von Grotthuss, P. Fontanillas, A. Spooner *et al.*, "Type 2 diabetes variants disrupt function of slc16a11 through two distinct mechanisms," *Cell*, vol. 170, no. 1, pp. 199–212, 2017.

[7] P. Sahlén, I. Abdullayev, D. Ramsköld, L. Matskova, N. Rilakovic, B. Lötstedt, T. J. Albert, J. Lundeberg, and R. Sandberg, "Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution," *Genome biology*, vol. 16, no. 1, p. 156, 2015.

[8] P. H. L. Krijger and W. De Laat, "Regulation of disease-associated gene expression in the 3d genome," *Nature reviews molecular cell biology*, vol. 17, no. 12, pp. 771–782, 2016.

[9] F. Zhang and J. R. Lupski, "Non-coding genetic variants in human disease," *Human molecular genetics*, vol. 24, no. R1, pp. R102–R110, 2015.

[10] S. C. Siu and C. K. Silversides, "Bicuspid aortic valve disease," *Journal of the American College of Cardiology*, vol. 55, no. 25, pp. 2789–2800, 2010.

[11] I. Mordi and N. Tzemos, "Bicuspid aortic valve disease: a comprehensive review," *Cardiology research and practice*, vol. 2012, 2012.

[12] A. Ameur, J. Dahlberg, P. Olason, F. Vezzi, R. Karlsson, M. Martin, J. Viklund, A. K. Kähäri, P. Lundin, H. Che *et al.*, "Swegen: a whole-genome data resource of genetic variability in a cross-section of the swedish population." *European journal of human genetics: EJHG*, 2017.

[13] SweGen. (2016 (accessed June 16, 2017)) Swedgenn variant frequency browser. [Online]. Available: https://swegen-exac.nbis.se/

[14] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry *et al.*, "The variant call format and vcftools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011.

[15] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, "A

framework for variation discovery and genotyping using next-generation DNA sequencing data." *Nature genetics*, vol. 43, no. 5, pp. 491–498, May 2011. [Online]. Available: http://dx.doi.org/10.1038/ng.806

[16] H. Li, "Toward better understanding of artifacts in variant calling from high-coverage samples," *Bioinformatics*, vol. 30, no. 20, pp. 2843–2851, 2014.

[17] T. Ringo Oki, S; Ohta. (2015 (accessed Sep20, 2017)) Chip-atlas. [Online]. Available: https://github.com/inutano/chip-atlas/

[18] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'donovan, and R. Apweiler, "Quickgo: a web-based tool for gene ontology searching," *Bioinformatics*, vol. 25, no. 22, pp. 3045–3046, 2009.

[19] A. Mora, G. K. Sandve, O. S. Gabrielsen, and R. Eskeland, "In the loop: promoter–enhancer interactions and bioinformatics," *Briefings in bioinformatics*, vol. 17, no. 6, pp. 980–995, 2015.

[20] J. C. Grisar, F. Haddad, F. A. Gomari, and J. C. Wu, "Endothelial progenitor cells in cardiovascular disease and chronic inflammation: from biomarker to therapeutic agent," *Biomarkers*, vol. 5, no. 6, pp. 731–744, 2011.