Knowledge-Constrained K-Medoids Clustering of Regulatory Rare Alleles for Burden Tests

R. Michael Sivley, Alexandra E. Fish, and William S. Bush

Center for Human Genetics Research, Department of Biomedical Informatics,
Vanderbilt University, Nashville, TN, USA
{mike.sivley,alexandra.e.fish,william.s.bush}@vanderbilt.edu

Abstract. Rarely occurring genetic variants are hypothesized to influence human diseases, but statistically associating these rare variants to disease is challenging due to a lack of statistical power in most feasibly sized datasets. Several statistical tests have been developed to either collapse multiple rare variants from a genomic region into a single variable (presence/absence) or to tally the number of rare alleles within a region, relating the burden of rare alleles to disease risk. Both these approaches, however, rely on user-specification of a genomic region to generate these collapsed or burden variables, usually an entire gene. Recent studies indicate that most risk variants for common diseases are found within regulatory regions, not genes. To capture the effect of rare alleles within non-genic regulatory regions for burden tests, we contrast a simple sliding window approach with a knowledge-guided k-medoids clustering method to group rare variants into statistically powerful, biologically meaningful windows. We apply these methods to detect genomic regions that alter expression of nearby genes.

1 Introduction

Numerous studies have been published illustrating the association of commonly occurring genetic variants to traits of interest in humans [1], and to changes in gene expression [2]. Recent technological advances in sequencing technology have enabled the study of rare variation – single base-pair changes in DNA that occur at less than 5% frequency in a population [3]. Typical genetic association studies rely on linear or logistic regression models to contrast the phenotype of interest across genotype categories based on a single variant (i.e. AA [25%], Aa [50%], and aa [25%]). Statistical power for these studies is directly related to the frequencies of these genotype categories, and lower frequency variants often have extremely low power to detect associations using these methods because most individuals in the study do not have the rare variant (i.e. AA [98%], Aa [1.8%], and aa [0.2%]).

Multiple methods have been proposed to address the issues of statistical power [4], all of which rely on grouping rare variants together either by biological function or physical proximity in the genome. The vast majority of these statistical methods provide users with the flexibility to specify the genomic region they wish to use for grouping variants together. In practice, variants are typically collapsed within gene

L. Vanneschi, W.S. Bush, and M. Giacobini (Eds.): EvoBIO 2013, LNCS 7833, pp. 35–42, 2013.

regions under the hypothesis that a variants influence disease by changing coding DNA that impacts protein function in some way. However, recent publications by the ENCODE project have shown that the vast majority of previously identified genetic associations are non-coding and regulatory in nature[5].

Currently, non-genic approaches to group rare variants include a simple sliding window approach [6] or collapsing variants within regions defined by experimental data, such as the ENCODE annotations. Sliding window approaches require millions of statistical tests which are highly correlated. The large number of tests makes determining the false positive or false discovery rate of the analysis challenging. Collapsing variants within putative regulatory regions may produce windows that are too small to capture variants to provide a powerful test. This approach also assumes that the genomic locations of regulatory regions are well-defined – an unlikely assumption for many Chromatin Immuno-Precipiation (ChIP) experiments [7]. Therefore, new methods for defining non-genic windows for statistical analysis are needed.

In this work, we apply k-medoids clustering to leverage both physical proximity and biological function with the goal of defining groups of rare variants for statistical analysis. We use a single source of putative biological function – a prediction of genome function based on chromatin state – and refine groupings using physical proximity in the genome. We apply this clustering method to generate rare variant groupings and evaluate the impact of these grouped variants on gene expression traits. Results from our clustering-based approach are compared with a traditional sliding window approach.

2 Methods

2.1 Data

Publically available datasets with phased haplotype information and whole-genome gene expression data on 1000 Genomes samples were used [3]. There were 149 independent, multi-ethnic individuals, consisting of 32 CEPH (CEU) and 37 Yoruba (YRI) parental samples, and 41 Chinese (CHB) and 39 Japanese (JPT) unrelated individuals. Phased haplotype data was obtained from the imputation reference panels for MaCH software (1000G Phase 1 version 3 MaCH panels) and was based upon 1000 Genomes Phase 1 integrated genotype calls and included singleton variants [8]. For gene expression data, we accessed normalized gene expression data from [2] (available online: http://eqtnminer.sourceforge.net/), which was generated using Illumina human whole-genome expression arrays (WG-6 version 1) on lymphoblastoid cell lines from each of the 149 individuals. Expression data was first normalized by quantile normalization within replicates, and then was median normalized across individuals. Additionally, we applied Gaussian quantile normalization for the test genes within each population, in order to account for population differences in gene expression. This normalization was congruent with the original normalization performed in [2]. For each of the selected genes, we extracted genotypes in the *cis*-regulatory region (500KB upstream of the transcriptional start site and 500KB downstream of the transcriptional end site).

2.2 Domain Knowledge

We used classification results from a published study of chromatin marks [9] to guide our cluster analysis. This study used ChIP data to identify methylation and acetylation modifications to histone proteins throughout the genome for nine cell lines. These patterns form the *histone code* [10], and were classified using a multivariate Hidden Markov Model into 15 states, which we loosely grouped into promoter, enhancer, insulator, and transcribed regions. Because our analysis was focused exclusively on gene expression in lymphoblastoid cell lines, we used chromatin state classifications generated for the GM12878 lymphoblastoid cell line. This data is available via the ENCODE project website through the UCSC genome browser (http://genome.ucsc.edu/ENCODE/). By guiding our cluster analysis with this data, we hypothesize that genetic variation within similar chromatin states should be grouped together.

2.3 Gene Selection

To compare the two methods across a variety of different regulatory architectures, four genes were selected from a group of genes previously identified as having collections of rare variants functioning as cis-eQTLs, based upon a genome-wide collapsing analysis (unpublished data). Each gene selected represents a potentially unique regulatory architecture, based upon the functional annotation of rare variants which were within the significant regions. Rare variants within significant regions could be identified as disrupting a transcription factor binding site (*ORMDL1*), being present in a ChIP peak (*NUDT22*), or having no functional annotation whatsoever (*FAM154B*). A potential confounder to this study is the presence of common eQTLs in significant regions. A compilation of known common eQTLs was used to determine that none of the above genes had a common eQTL in the previously identified significant regions. To interrogate the effects of common eQTLs on the analysis, *DYPSL4* was also selected, which contained three common eQTLs in the previously identified significant region in addition to rare variants affecting transcription factor binding sites.

2.4 Cluster-Based Analysis

Constrained Partitioning (COP) is a method by which partial knowledge can be introduced into a clustering algorithm, making it a semi-supervised method. Constraints allow for otherwise uninformed clustering methods to include background knowledge of a particular domain. Typically, COP is provided with a list of must-link constraints and cannot-link constraints, which dictate which observations must and cannot be placed in the same cluster.

In our implementation, we allow for an initial classification of chromatin state SNPs surrounding a gene. This classification acts as a must-link constraint for all observations in a class, and a cannot-link constraint for all observations of differing classes. We then apply Partitioning Around Medoids (PAM) to subdivide these SNPs

according to their base position. PAM divides the data into k clusters, where k is specified *a priori* [11]. To choose an optimal k, we ran PAM multiple times with increasing k and select k such that it maximizes with average silhouette width of the resultant clusters. The choice of k is made for each initial classification and the original classes do not need to be partitioned into the same number of clusters.

With our rare variants clustered, we then performed a rare variant burden test, which collapses the data into a single variable, indicating the number of rare variants within that cluster. For each cluster, linear regression was used to determine the significance of association between the clustered rare variants and gene expression. This implementation was done entirely in R.

2.5 Sliding Window Analysis

A rare variant burden test with sliding windows was performed on the test genes. For each gene, the region tested consisted of 500KB both up and downstream, in addition to the gene itself. In this region, a 5KB sliding window was used, such that each SNP served as the start point for a window. All rare variants in this 5KB region were used to determine the burden of rare variants. Only windows with at least one rare variant detected were included in analysis. For each window, a linear regression was performed between the number of rare variants present within a region for each individual and the gene expression level. This is slightly different from the analysis used to select the genes, in which individuals were placed into a binary category of either having a rare variant or not – a *collapsing* test [12].

2.6 Determination of Significance

The best practice for the statistical analysis of sliding windows is a current topic of debate. To place these results in the context of standard genetic analysis guidelines, both a Bonferroni correction and a False Discovery Rate (FDR) analysis were performed [13]. Each gene was analyzed independently in both the Bonferroni and FDR (FDR = 0.05) analyses. In the Bonferroni correction analysis, the number of clusters present in each gene is used to set the gene-specific significance threshold for cluster data. For the sliding window analysis, the number of windows set the gene-specific significance threshold. After being identified as significant, all overlapping windows were merged to form a significant 'signal' in the sliding window analysis.

2.7 Visualization

We visualized the results from both the sliding window and cluster analyses in a single plot using the R package ggplot2 [14]. For the sliding window analysis, the midpoint chromosome position of each 5KB window is plotted relative to the $-\log_{10}$ of the regression p-value to generate a *Manhattan* plot. We used loess to fit a smooth curve to these data points using the stat_smooth function with a span parameter of

0.2. Results from the cluster analysis are shown as horizontal bars (to illustrate the span of the cluster) plotted relative to the $-\log_{10}$ of the regression p-value, color coded by chromatin state. Note that some clusters are too small to be seen on these plots.

3 Results

3.1 Gene Region Results

Visual comparisons of sliding window and cluster analysis approaches are provided in figure 1. *ORMDL1* best illustrates the potential of this method. A highly significant effect is seen from an enhancer cluster which overlaps with the strongest effect from the sliding window analysis. *NUDT22* also shows a strong effect of a large enhancer cluster which spans the best sliding window effect. For both these genes the clustering results correlate well with the loess curves, capturing the 'shape' of the regional effect. The cluster analysis shows less utility for *DYPSL4*, a gene with complex common eQTL effects, and *FAM154B,* a gene with no obvious regulatory mechanisms. For these genes, the method clustered together distant variants within insulator elements creating single clusters containing variants at great distances; these clusters do not reflect the domain knowledge well. We plan to refine the algorithm to include additional constraints limiting the physical distance separating rare variants within potential clusters.

3.2 Bonferroni Correction

The summary of significant genomic regions with a Bonferroni corrected analysis is presented in Table 1. Similar numbers of significant genomic regions are returned by both the sliding window and clustering analysis. In both methods, *DYPSL4* failed to result in significant results. In the case of *ORMDL1*, both clustering and sliding window analysis each resulted in one unique significant region which was not overlapping. All other significant regions overlapped with a region identified in the other test. In *NUDT22*, all significant signals identified by sliding window analysis overlapped with significant clusters. Cluster analysis additionally resulted in two unique significant regions. None of the significant regions identified in *FAM154B* overlapped between the sliding window analysis and the clustering analysis.

Table 1. Number of significant genomic regions detected using both clustering and sliding window analysis with a Bonferroni correction for multiple testing

GENE	Bonferroni Threshold for Cluster Analysis	Number of Significant Clusters	Bonferroni Threshold for Sliding Window Analysis	Number of Significant Windows from Sliding Window Analysis
ORMDL1	0.001250	6 of 40	3.95476×10^{-6}	604 of 12,643
NUDT22	0.001351	5 of 37	4.64857×10^{-6}	26 of 10,756
DYPSL4	0.001282	0 of 39	3.16476×10^{-6}	0 of 15,799
FAM154B	0.001351	3 of 37	6.38162×10^{-6}	32 of 7,835

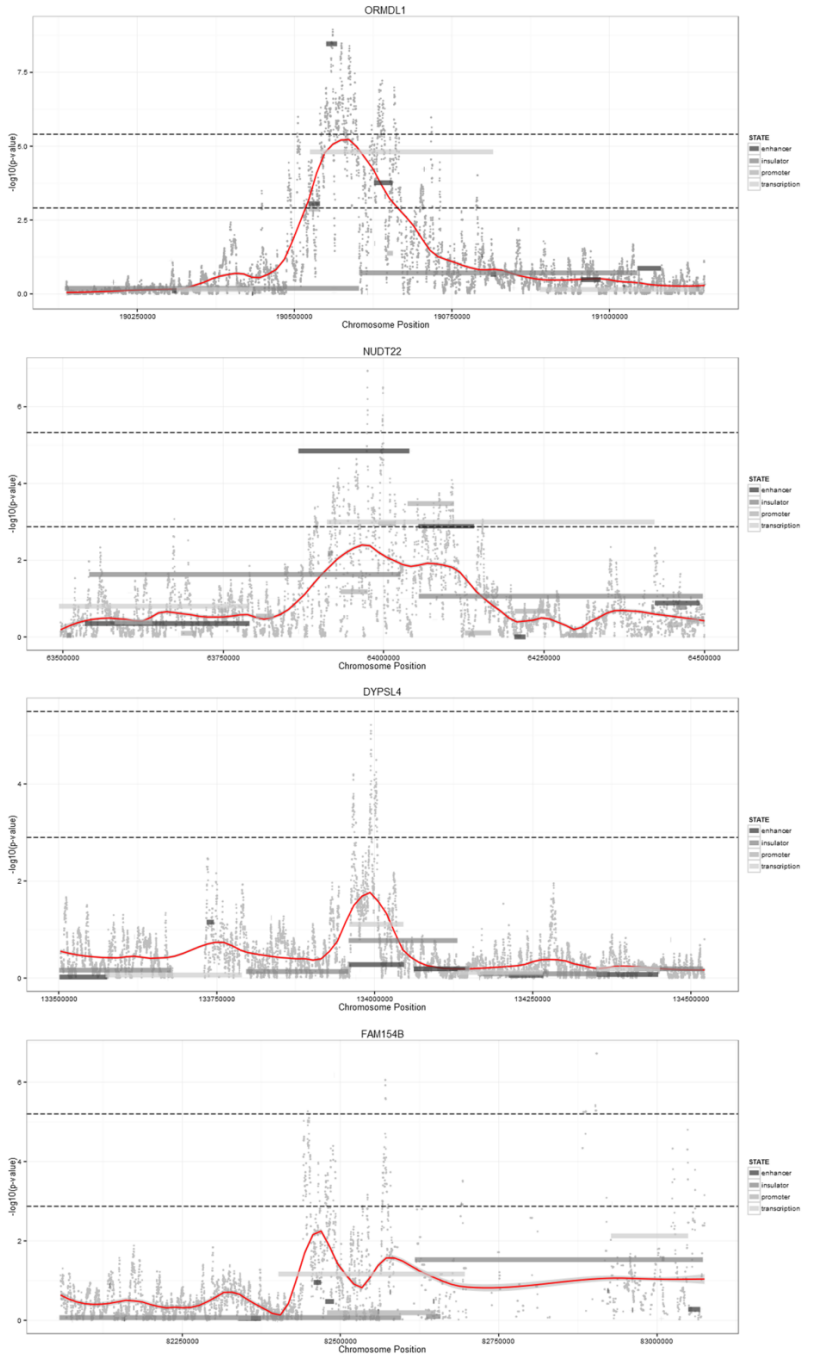

Fig. 1. Manhattan plot of window midpoints (points), variant clusters (bars) by significance with loess fit (red line, loess span = 0.2) of window midpoint by significance

3.3 False Discovery Rate Correction

The significant genomic regions with a FDR (FDR = 0.05) corrected analysis are presented in Table 2. All the regions identified as significant with the Bonferroni correction were identified with the FDR correction as well. One unique cluster was identified with FDR analysis in both *ORMDL1* and *NUDT22*. A dramatic increase was observed in the number of signals identified as significant in the sliding window analysis. For *ORMDL1*, *NUDT22*, and *FAM154B*, all significant clusters overlapped with regions identified as being significant by sliding window analysis. In the case of *DYPSL4*, clustering failed to identify any significant regions, whereas sliding window analysis identified two genomic regions as significant. Sliding window analysis identified a total of 28 unique genomic regions as significant in these genes.

Table 2. Number of significant genomic regions detected using both clustering and sliding window analysis with an FDR=0.05 correction for multiple testing. *There are no p-values < 0.05, making it impossible to calculate the FDR = 0.05 threshold.

GENE	Threshold for Cluster Analysis FDR = 0.05	Number of Significant Clusters	FDR = 0.05 Threshold for Sliding Window Analysis	Number of Significant Windows from Sliding Window Analysis
ORMDL1	0.001346812	7 of 40	0.007989149	2021 of 12,643
NUDT22	0.006583255	6 of 37	0.007619227	1126 of 10,756
DYPSL4	NA*	0 of 39	0.000434797	126 of 15,799
FAM154B	0.001213077	3 of 37	0.006232502	628of 7,835

4 Discussion

Our results indicate that informed clustering of rare variants using regulatory annotations can dramatically reduce the number of statistical tests, reducing the multiple testing burden for rare variant analysis, thus increasing overall power. Obviously, this approach will perform best when the underlying assumption of the method holds true; that influential variants fall within regulatory regions, as illustrated in the *ORMDL1* gene.

A great strength of this approach is that the clustering is independent of statistical analysis, and can be coupled with various methods, such as the Sequence Kernel Association Test(SKAT) or KBAC [15, 16]. Because the method is unsupervised, there are no over-fitting concerns in the association analysis, and standard statistical assumptions of these tests are not violated. The cluster method could also be informed by statistical power calculations of the coupled association test (or other testing assumptions), allowing clusters of rare variants to be optimized to improve the overall power of the analysis. Finally, in this study we have used chromatin state data to guide cluster formation, however numerous other genomic annotations could be applied simultaneously to intelligently design functional clusters of rare variants. As ENCODE and other projects continue to expand our understanding of gene regulation, methods that can leverage this data for analysis will become ever more important.

Acknoweldgements. This work was supported in part by NIH U01 HG004798 and its ARRA supplements.

References

1. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., Manolio, T.A.: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106, 9362–9367 (2009)
2. Veyrieras, J.-B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., Pritchard, J.K.: High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genetics 4, e1000214 (2008)
3. Durbin, R.M., Altshuler, D.L., Abecasis, G.R., Bentley, D.R., Chakravarti, A., et al.: A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010)
4. Bansal, V., Libiger, O., Torkamani, A., Schork, N.J.: Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics 11, 773–785 (2010)
5. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., Snyder, M.: Linking disease associations with regulatory information in the human genome. Genome Research 22, 1748–1759 (2012)
6. Lawrence, R., Day-Williams, A.G., Elliott, K.S., Morris, A.P., Zeggini, E.: CCRaVAT and QuTie-enabling analysis of rare variants in large-scale case control and quantitative trait association studies. BMC Bioinformatics 11, 527 (2010)
7. Mendenhall, E.M., Bernstein, B.E.: DNA-protein interactions in high definition. Genome Biology 13, 139 (2012)
8. Li, Y., Willer, C.J., Ding, J., Scheet, P., Abecasis, G.R.: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiology 34, 816–834 (2010)
9. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., Bernstein, B.E.: Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43–49 (2011)
10. Rando, O.J.: Combinatorial complexity in chromatin structure and function: revisiting the histone code. Current Opinion in Genetics & Development 22, 148–155 (2012)
11. Kaufman, L., Rousseeuw, P.: Clustering by means of medoids (1987)
12. Li, B., Leal, S.M.: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. American Journal of Human Genetics 83, 311–321 (2008)
13. Storey, J.D., Tibshirani, R.: Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 100, 9440–9445 (2003)
14. Wickham, H.: ggplot2: elegant graphics for data analysis. Springer, New York (2009)
15. Liu, D.J., Leal, S.M.: A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genetics 6, e1001156 (2010)
16. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X.: Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics 89, 82–93 (2011)