Published in final edited form as:
Pac Symp Biocomput. 2011 ; : 76–87.

AN APPLICATION AND EMPIRICAL COMPARISON OF STATISTICAL ANALYSIS METHODS FOR ASSOCIATING RARE VARIANTS TO A COMPLEX PHENOTYPE

VIKAS BANSAL[*],
The Scripps Translational Science Institute (VB, OL, AT) and Department of Molecular and Experimental Medicine, The Scripps Research Institute (AT); 3344 North Torrey Pines Court, Suite 300, La Jolla, CA 92037 USA

ONDREJ LIBIGER[*],
The Scripps Translational Science Institute (VB, OL, AT) and Department of Molecular and Experimental Medicine, The Scripps Research Institute (AT); 3344 North Torrey Pines Court, Suite 300, La Jolla, CA 92037 USA

ALI TORKAMANI[*], and
The Scripps Translational Science Institute (VB, OL, AT) and Department of Molecular and Experimental Medicine, The Scripps Research Institute (AT); 3344 North Torrey Pines Court, Suite 300, La Jolla, CA 92037 USA

NICHOLAS J. SCHORK[†]
The Scripps Translational Science Institute and Department of Molecular and Experimental Medicine, The Scripps Research Institute; 3344 North Torrey Pines Court, Suite 300, La Jolla, CA 92037 USA

VIKAS BANSAL: vbansal@scripps.edu; ONDREJ LIBIGER: olibiger@scripps.edu; ALI TORKAMANI: atorkama@scripps.edu; NICHOLAS J. SCHORK: nschork@scripps.edu

Abstract

The contribution of collections of rare sequence variations (or 'variants') to phenotypic expression has begun to receive considerable attention within the biomedical research community. However, the best way to capture the effects of rare variants in relevant statistical analysis models is an open question. In this paper we describe the application of a number of statistical methods for testing associations between rare variants in two genes to obesity. We consider the relative merits of the different methods as well as important implementation details, such as the leveraging of genomic annotations and determining p-values.

[†] To whom correspondence should be addressed: nschork@scripps.edu.
[*] VB, OL and AT contributed equally to this paper.

1. Introduction

1.1. Rare variants and the ‘hidden heritability’ of complex traits

Genome wide association (GWA) studies have been pursued for many diseases and phenotypes. Although the results of these studies have been mixed, with some studies identifying more compelling associations than others, virtually all of these studies have resulted in the discovery of variants that collectively only explain a small fraction of the heritable component of the diseases and phenotypes they have considered [1]. This fact has not only raised important questions about the degree to which common variants, which are typically of focus in GWA studies, influence phenotypic expression, but also the best way to identify factors not detectable via current common-variant-based GWA study protocols [2] that contribute to a ‘hidden heritability’ behind phenotypic expression.

Recently it has been argued that collections of rare variants could contribute to phenotypic expression over-and-above common variants [3–4]. The intuition behind this argument is that although each rare variant may have a small overall effect on phenotypic expression, collectively these variants may have a moderate or even more pronounced effect [3–4]. Rapid developments in high-throughput DNA sequencing technologies are likely to facilitate searches for rare variants that may influence phenotypic expression, but are not the only item necessary for a successful study of rare variants. Also needed are appropriate study designs and subject sampling methods, data analysis methods, and ways of validating or conceptualizing the biological influence of multiple rare variants on phenotypic expression once they are found to be associated with a phenotype.

In this paper we describe a number of different statistical methods for testing the hypothesis that collections of rare variants are associated with a qualitative phenotype in a case/control sampling setting. These methods build off the notion of ‘collapsing’ a number of rare variants into a single set whose collective frequency is contrasted between case and control groups [5–6]. Many approaches involve regression or regression-like models in which dummy variables indicating the presence (i.e., individuals assigned a dummy variable value of 1.0) or absence (0.0) of a variant are used. For the collapsed set of variants, an individual is ultimately assigned a value of 1.0 if they have any of the rare variants among a larger set and 0.0 otherwise. This collapsed dummy variable can then be tested for association by testing the regression coefficient associated with the dummy variable [7]. Other regression approaches consider the effects of each individual variant, no matter how rare, as well as collapsed sets of variants [8]. We apply these and other methods to a case/control study of obesity and compare the results of the application of each. We also consider extensions of the proposed statistical analysis methods.

Before describing the data set, statistical methods, and the results of their application, however, we provide brief descriptions of two overarching frameworks for the study of the collective effects of rare variants on phenotypic expression: one leveraging functional genomic annotations and one considering the collective effects of variants in defined contiguous genomic regions.

1.2. Collapsing variants based on functional annotations

Testing collections of rare variants for association to a phenotype requires some way of grouping or collapsing variants into a coherent set; i.e., defining the set whose collective frequency is tested for association. This can be approached by defining a set based on functional annotations associated with the genomic regions harboring the variants to be tested for association. For example, one could test the collective frequency differences of coding variants, non-synonymous coding variants, variants in known transcription factor binding sites, or conserved sites, between cases and controls. Such groupings could lead to easily interpreted biological associations but, ultimately, would only be as good and reliable as the annotations used.

1.3. Moving window analysis

An alternative to defining sets of collapsed variants based on functional genomic annotations is to consider all the variants in a genomic subregion defined by its size and test these variants for association. Such subregions could then be systematically tested over the entire genomic region of interest. By starting at one end of a genomic region of interest, testing variants within the ‘window’ defined by the subregion, and then moving the window to an adjacent subregion, testing that subregion, and continuing this process until the entire region is covered would provide a test of the hypothesis that some subregions within the broader region of interest harbor collections of variants associated with a phenotype. This moving window approach can be repeated with different window sizes, including overlapping windows, but at the cost of increased type I error due to the multiple tests.

1.4. Accommodating other sources of variation and assessing statistical significance

In any test of genetic association there are a few things that need to be considered. For example, stratification issues need to be accommodated or controlled for. This can be done by ensuring that the subjects used in a study are matched for genetic background or the statistical test used is appropriately adjusted for potential stratification [9]. In addition, in order to assess the statistical significance of an association study involving multiple variants within a genomic region, appropriate control for multiple comparisons must be made [10]. Finally, accommodating covariate effects (e.g., gender, age, other genetic factors, ancestry information, etc.) in association analysis is important, but may not be trivial for many statistical models. Thus, gauging the ability of different statistical analysis models to accommodate covariates may be of particular importance in rare variant analysis settings.

2. Sequencing the MGLL and FAAH genes in obese and control individuals

2.1. DNA sequencing and sample selection

Genomic intervals covering two genes that encode the endocannabinoid metabolic enzymes, FAAH and MGLL, were sequenced in 289 individuals of European ancestry using the Illumina GA sequencer. Ancestry was determined using a panel of ancestry informative markers and individuals with an outlying genetic background were removed from the analysis. Sequencing was done using 36 base pair reads. The median coverage was 60X across the individuals sequenced. The program MAQ was used for alignment and variant calling, resulting in 1410 high quality single nucleotide variants (SNVs; 228 in the FAAH gene and 1182 in the MGLL gene) which were used for association analysis. The sequenced regions were captured using long range PCR and represented a total of 188,270 nucleotides. The 289 individuals included 147 normal controls (Body Mass Index (BMI) <30) and 142 extremely obese cases (BMI >40).

2.2. Genomic annotations, window definitions, and multiple comparisons

We leveraged genome annotations from the UCSC genome browser to identify sets of variants that reside in functionally-relevant regions of the genome. We identified sets of variants that reside within 5 different functional elements within the MGLL and FAAH genes: non synonymous SNVs (‘NS’), H3K27 acetylation sites, Fox2 interaction sites, Amidase protein domains, and all transcription factor binding sites (‘TFBS’). Variants within these elements were collapsed and tested for association with obesity. For the moving window analyses, we considered window sizes of 5 kb over the two genes. In order to accommodate multiple comparisons we identified the effective number of independent variants based on linkage disequilibrium (LD) using the method discussed by Nyholt [11]. This number provides a very rough approximation for the number of tests to be corrected for and was found to be 584 for our data. We assumed a nominal type I error rate of 0.05 in assessing statistical significance of variant associations, so our approximate multiple comparisons corrected p-value was 0.05/584 = 0.000086 (−log(p-value) = 4.06). Obviously, more sophisticated strategies for correcting for multiple comparisons, including possibly permuting cases and controls and repeating the entire moving window and functional annotation-based collapsed set analyses, need to be investigated.

3. Statistical methods for rare variant associations

We briefly describe 11 methods that can be used to test the hypothesis that collections of rare variants are associated with a phenotype. We also consider 9 high-dimensional regression and data mining procedures that can be used to simultaneously test the association of all individual variants, rare and common, as well as collapsed sets of variants. We did not consider covariates in these analyses. Space limitations preclude an in-depth discussion of each method so we provide references and only the main intuitions behind each method.

3.1. Single locus and general collapsed variant test-based methods

The following very brief descriptions of the methods we considered. Many of the papers describing these methods include discussions of possible extensions or alternative formulations of each method. We chose what we believe is the strategy that best represents the approaches described in those papers.

Single-locus tests (SL)—We considered the use of Fisher’s exact test to assess the association between each SNV and morbid obesity case/control status. We pursued single locus tests as a contrast for the multilocus-based collapsed variant tests since the power to detect an association involving a rare SNV is low.

Li and Leal Collapsing Method (LL)—Li and Leal [6] proposed a collapsing method for testing for association with multiple rare variants. Briefly, the method collapses the genotype information across multiple (rare) variants into a single variable for each individual. This new variable can then be tested for association with a phenotype using a chi-square test or the Fisher exact test. Given a collection of variants (grouped together based on function or position in a genomic region), we considered the subset of variants with a low minor allele frequency (MAF <0.02). Additionally, variants with virtually no difference in allele frequency between the cases and controls (Fisher test p-value >=0.6) were also removed. Using the remaining variants, a binary variable was defined for each individual as 1 if the individual had the rare allele for any of the variants and 0 otherwise. Fisher's exact test was used to compute the significance of the difference in allele frequency of this binary variable between the cases and controls. The p-value of the statistic was computed by permuting the case-control status of the individuals and determining the fraction of permutations for which the statistic was lower than or equal to the observed statistic.

Madsen and Browning Method (MB)—We implemented the groupwise association test described by Madsen and Browning [12]. Given a group of variants, this method tests for the presence of an excess of rare SNVs in the cases as compared to the controls. Each SNV is given a weight based on its minor allele frequency in the controls. A score is calculated for each individual using the individual genotypes and the weights of each variant. The sum of ranks of scores of the cases is used as the statistic (similar to Wilcoxon rank test). We computed the p-value for each statistic using a maximum of 1000 permutations. The test was performed using the 'general disease' model described by Madsen and Browning [12]. This model only allows for the analysis of rare variants and does not accommodate the effects of common variants.

Subset Selection Method (SS)—Recently, Bhatia et al. [13] have proposed an extension of the Collapsing method of Li and Leal. Instead of collapsing across all rare variants in a set, the method searches for a subset of variants which maximally discriminate between the cases and controls. The method described by Bhatia et al. [13] uses a greedy algorithm to identify a subset of variants for which their collective occurrence or union has a large difference in frequency between the case and control individuals. This model only allows for the analysis of rare variations (MAF < 0.02) and does not accommodate the effects of common variations. Fisher's exact test was used to assess the significance of sets of variants at any point in the search for the optimal set.

Distance-based diversity (Dis)—Distances between the diploid sequences of all pairs of individuals in the study were calculated as one minus the identity-by-state similarity across the variant loci in a set. The dispersion of (i.e., variation among) the sequences within and between case and control groups was then compared using the 'betadisper' function of the 'vegan' package (version 1.17–0) in the R computing environment [14]. This function essentially implements Anderson's [15] PERMDISP2 procedure for the analysis of multivariate homogeneity of group dispersions [15]. Tests of the hypothesis that there is greater diversity among the cases or controls was assessed empirically via a permutation test implemented in the function 'permutest' in the PERMDISP2 package.

Omnibus haplotype frequency test (PHap)—We considered the omnibus haplotype test strategy outlined by Fallin et al. [16] and Zhao et al. [17] and implemented in PLINK [18] for sets of variants in contiguous regions. This approach essentially tests the hypothesis that haplotype frequency profiles are equal between cases and controls, where the haplotypes harbor the variants of interest.

Power-based diversity statistic Gst (Div)—We tested the hypothesis that for any set of variants the cases and controls would differ in terms of the diversity they exhibited across those variants. To conduct an appropriate test of this hypothesis, we implemented the procedure for assessing population differentiation based on the measure Gst described in equation 8 of Jost [19].

Sequence similarity statistic leveraging MDMR (Sim)—We considered the use of the multivariate distance matrix regression (MDMR) and Generalized Analysis of Molecular Variance (GAMOVA) approaches discussed by Wessel et al. [20] and Nievergelt et al. [21] to test the hypothesis that the multilocus genotype profiles encompassing a set of variant loci exhibited by the cases were more similar amongst themselves than with the controls. Distances between pairs of sequences were calculated by subtracting the average value of identity-by-state similarity across loci in each window from one. The approach was implemented by O. Libiger and M. Zapala in Python (script available at http://polymorphism.scripps.edu/~cabney/). Permutation tests were used to assess statistical significance of any differences in similarity.

Ridge regression (Ridge)—We used ridge regression to test the hypothesis that individual variants and collapsed sets of variants (made into a dummy variable, as described in section 1.1 above) were associated with obesity level. We used the approach outlined by Malo et al. [22] for this analysis. The method of Hoerl, Kennard, and Baldwin [23] was used to estimate the ridge parameter.

Logic regression (Logic)—We also considered logic regression to identify combinations of variants that were associated with obesity. We used the implementation of logic regression that is available in the R computing environment package ‘LogReg’ [24]. We fit two logic trees and performed a null-model permutation test to assess significance of the association between identified sets of variants and case/control status.

Set based analysis (PSet)—We considered variant set-based tests similar in orientation to Fisher’s combined p-values methodology [25]. We use the method implemented in the PLINK software package for this analysis [18]. PLINK default parameters were used throughout the analysis. Statistical significance was assessed via a permutation test.

3.2. High-dimensional regression methods

As noted, we also considered the analysis of the data using high-dimensional regression and data mining procedures. These procedures could essentially consider all the variants, both in isolation or in collapsed sets, as predictors of the phenotype and were not used in moving window analyses.

Lasso (L)—We considered the use of Lasso-based regression [26] using 'bridge' regression with the penalty parameter set to 1.0 and all other parameters set to their default value [27], as implemented in the 'gpsbridge' function of the R/GPS interface developed by Jerome Friedman for the R computing environment [14]. 10-fold cross validation was performed to select the best model.

Generalized path seeking regression (GPS)—We employed 'bridge' regression with all parameters set to their default value [27]), as implemented in the 'gpsbridge' function of the R/GPS interface developed by Jerome Friedman for the R computing environment [14]. 10-fold cross validation was performed to select the optimal model and penalty value.

Stepwise Regression (SR)—We performed stepwise linear model selection via the Akaike Information Criterion (AIC) for choosing associated variants and collapsed variant sets using the function 'stepAIC' from the package 'MASS' developed for the R computing environment [14].

Classification and regression trees (SPM-CART)—We considered the CART method originally described by Breiman et al. [28] and implemented in the Salford systems data mining software suite (http://salford-systems.com/) to identify predictors of obesity.

Multiple adaptive regression trees (SPM-TreeNet)—We also used the TreeNet procedure originaly described by Friedman et al. [29] and implemented in the Salford systems data mining software suite (http://salford-systems.com/).

Multivariate adaptive regression splines (SPM-MARS)—We implemented the MARS procedure originally developed by Friedman [30] and implemented in the Salford systems data mining software suite (http://salford-systems.com/).

Random Forests (SPM-RF)—We explored the use of the Random Forests procedure introduced by Breiman [31] and implemented in the Salford systems data mining software suite (http://salford-systems.com/).

Conjunctive rule learner (Weka CRL)—We considered the conjunctive rule learner algorithm as described by Witten and Frank [32] and implemented in Weka [33] with no ranking.

Representative tree (Weka REPTree)—We used the representative tree algorithm as described by Witten and Frank [32] and implemented in Weka [33].

4. Results

4.1. Collapsed variants based on functional annotations

We first considered the significance of the difference of variants within the five functional elements derived from annotations for the FAAH and MGLL genes discussed in section 2.2. Table 4.1 provides the p-values associated with 10 multilocus data analysis methods described in section 3.1 (we did not consider single locus analyses here). From Table 4.1 it can be seen that, with the exception of an analysis of collapsed variants within all transcription factor binding sites (the 'TFBS' column in Table 4.1) for the MGLL gene, there is not consistent evidence for association among the different methods.

4.2. Moving window analysis

We considered the application of the 11 different analysis methods to a moving window analysis of the MGLL and FAAH genes. The analysis explored adjacent windows of size 5000 bases for the both the MGLL and FAAH genes. The −log(p-value) computed for each test is plotted on the y-axis of Figure 4.2 against the midpoint of each window. The different panels (i.e., contour plots) reflect different analysis methods, which are, from bottom to top: standard single locus analysis using Fisher's exact test (SL); Li and Leal's [6] method (LL); the Madsen and Browning [12] weighted average statistic (MB); the optimal subset selection method [13]; SS); the sequence distance-based diversity statistic based on the method of Anderson [15] (Dis); the sequence diversity statistic based on the power statistic of Jost [19] (Div); the sequence similarity based statistic discussed by Wessel et al. [20] and Nievergelt et al. [21] (Sim); the ridge regression statistic [22]) (Ridge); the Logic Regression [24]) statistic (Logic); the omnibus haplotype frequency test implemented in the PLINK software package [16–18] (Phap); and the set based analysis method implemented in the PLINK software package [18] (Pset). As noted in section 2.2, a −log(p-value) of 4.06 provide some correction for an overall multiple comparisons type I error rate of 0.05. It does not appear that any of the windows produces a −log(p-value) that would be significant after multiple comparisons corrections. In addition, many of the contour plots do not appear to track together, suggesting that the various data analysis methods do not produce correlated test statistics or evidence for association. Although there is some suggestion of consistency of a signal in the 'rightmost' region of the MGLL gene, its significance is debatable. Similar conclusions were drawn from the analysis of the FAAH gene (data not shown).

4.3. Correlations between statistics

We assessed the correlations between the test statistics obtained over the moving window analyses of the two genes. We did not include single locus analyses or the set (Pset) and haplotype analysis (Phap) methods implemented in PLINK as part of this analysis. This provides some indication as to whether or not the different statistical methods are capturing the same signals. Table 4.3 provides the Spearman non-parametric correlation coefficients between the test statistics computed over the windows.

The shaded cells within Table 4.3 reflect significant correlations ($p<0.05$). It should be recognized that the majority of test statistics computed in the window-based analyses are not themselves statistically significant. Therefore, the value of the test statistics that went into the calculation of the correlations may reflect noise which clearly will affect the correlation strength between the test statistics. Despite this, some of the test statistics do exhibit correlations and therefore may be essentially capturing the same types of collective effects. For example, Ridge and Logic regression are highly correlated, as are the subset selection (SS) and Li and Leal's [6]; (LL) method. Many methods are not correlated, suggesting that they may either suffer from flaws, have low power, or are more powerful to detect different types of effects. Obviously, simulation studies could be used to sort this out.

4.4. High-dimensional regression analysis

We also considered the use of the nine high-dimensional regression and data mining procedures listed in section 3.2, as well as the ridge regression procedure discussed in section 3.1, to simultaneously evaluate the association of each SNV, rare and common, in addition to the 10 collapsed sets of variants within each of the two genes described in sections 1.2 and 4.1 to obesity. The various procedures tested are designed to identify the minimal set of factors that are predictive of a dependent variable and hence may have an ability to capture or identify variants causally associated with obesity. Table 4.4 lists the five most significant factors identified from the 10 different procedures in addition to providing the adjusted R-squared and the root mean squared error characterizing the fit of the model that includes those 5 factors. Note that individual SNVs are denoted by a number (e.g., 166) and the gene within which they reside (MGLL or FAAH) whereas collapsed sets of variants are denoted by their labels as defined in section 1.2. From Table 4.4 it can be seen that although some factors appear in the list of five factors for different methods (e.g., individual SNV 166 appears on the list for ridge regression (RR), the Lasso (L), and the GPS method, most of the factors identified for any method are unique to that method or just a few of the methods. This suggests that the different methods are likely to disagree about which factors are the most strongly associated with a phenotype. This may be a function of the purpose and design of these methods, which is for making reliable predictions and not necessarily detecting the strongest associations among a large set of potential predictors.

5. Conclusions and Future Directions

Studies investigating the role of rare variants in phenotypic expression and disease susceptibility will be pursued routinely in the not-so-distant future as sequencing technologies improve in efficiency. The ability to exploit these technologies will depend critically on an ability to assemble and organize sequence data as well as an ability to draw reliable inferences concerning the statistical (and biological) significance of differences in combinations of sequence variants between individuals with and without a particular phenotype. We have considered a number of different approaches for relating collections of rare sequence variants to a phenotype. We compared these methods on actual sequence data obtained from two genes in a study of morbidly obese and control subjects. Some of these methods (e.g., Logic, MDMR, Dis) are computationally intensive, which may complicate their utility in very large studies. Although we did not find overwhelming evidence for an association with obesity, our studies suggest that different analysis methods, not surprisingly, do not necessarily agree on the strength of associations.

This raises important questions as to why this is so and whether or not some statistical methods may be more powerful for detecting certain types of association over other approaches. In addition, if it is the case that one or another of the proposed methods is better at picking up a certain type of association signal (e.g., most methods are likely to be better for detecting multiple independent acting variants whereas a few, such as similarity based methods [20], may be better at detecting synergistically-acting variants) then a researcher might consider analyzing their data with different analysis methods and possibly different window sizes. This in turn raises questions about false positive rates due to the use of multiple analysis methods and the pursuit of multiple comparisons. In addition, the robustness of the methods to outliers, their level accuracy, ultimate power in various settings, and their ability to accommodate covariates all need to be explored. Many of these questions can be addressed by exploring both the theoretical derivation of different methods as well as their behavior in contrived, simulated data settings [34]. Such activity will be crucial if progress is to be made in understanding the contribution of rare variants to the genetic basis of complex phenotypes.

Acknowledgments

This work was supported in part by the following research grants: U19 AG023122-05, R01 MH078151-03, N01 MH22005, U01 DA024417-01, P50 MH081755-01, R01 AG030474-02, N01 MH022005, R01 HL089655-02, R01 MH080134-03, U54 CA143906-01, UL1 RR025774-03], the Price Foundation and Scripps Genomic Medicine. Ondrej Libiger is also supported by a grant from Charles University (GAUK number 134609). The Obesity data were generated as a result of a collaboration between the Scripps Translational Science Institute and Sanofi-Aventis.

References

1. Manolio TA, Brooks LD, Collins FS. J Clin Invest. 2008; 118:1590. [PubMed: 18451988]
2. Manolio TA, et al. Nature. 2009; 461:747. [PubMed: 19812666]
3. Bodmer W, Bonilla C. Nat Genet. 2008; 40:695. [PubMed: 18509313]
4. Schork NJ, Murray SS, Frazer KA, Topol EJ. Curr Opin Genet Dev. 2009; 19:212. [PubMed: 19481926]
5. Morgenthaler S, Thilly WG. Mutat Res. 2007; 615:28. [PubMed: 17101154]
6. Li B, Leal SM. Am J Hum Genet. 2008; 83:311. [PubMed: 18691683]
7. Morris AP, Zeggini E. Genet Epidemiol. 2010; 34:188. [PubMed: 19810025]
8. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. Plos Genetics. 2008; 4
9. Price AL, et al. Nat Genet. 2006; 38:904. [PubMed: 16862161]
10. Storey, JD. International Encyclopedia of Statistical Science. Lovric, M., editor. 2010.
11. Nyholt DR. Am J Hum Genet. 2004; 74:765. [PubMed: 14997420]
12. Madsen BE, Browning SR. PLoS Genet. 2009; 5:e1000384. [PubMed: 19214210]
13. Bhatia G, et al. PLoS Computational Biology. 2010 in press.
14. Venables, WN.; Ripley, BD. Statistics and computing. 4. Springer; New York: 2002. Modern applied statistics with S; p. xip. 495
15. Anderson MJ. Biometrics. 2006; 62:245. [PubMed: 16542252]
16. Fallin D, et al. Genome Res. 2001; 11:143. [PubMed: 11156623]
17. Zhao JH, Curtis D, Sham PC. Hum Hered. 2000; 50:133. [PubMed: 10799972]
18. Purcell S, et al. Am J Hum Genet. 2007; 81:559. [PubMed: 17701901]
19. Jost L. Mol Ecol. 2008; 17:4015. [PubMed: 19238703]
20. Wessel J, Schork NJ. Am J Hum Genet. 2006; 79:792. [PubMed: 17033957]
21. Nievergelt CM, Libiger O, Schork NJ. PLoS Genet. 2007; 3:e51. [PubMed: 17411342]
22. Malo N, Libiger O, Schork NJ. American Journal of Human Genetics. 2008; 82:375. [PubMed: 18252218]
23. Hoerl E, Kennard RW, Baldwin KF. Communications in Statistic - Simulation and Computation. 1975; 4:105.
24. Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L. Genet Epidemiol. 21(Suppl 1):2001.
25. Hoh J, Ott J. Nat Rev Genet. 2003; 4:701. [PubMed: 12951571]
26. Tibshirani R. Journal of the Royal Statistical Society Series B-Methodological. 1996; 58:267.
27. Friedman JH. Fast sparse regression and classification. Stanford University Technical Report. 2008

28. Breimann, L.; Friedmann, JH.; Olshen, RA.; Stone, CJ. Wadsworth, editor. Pacific Grove; 1984. p. 385

29. Friedman J, Hastie T, Tibshirani R. Annals of Statistics. 2000; 28:337.

30. Friedman JH. Annals of Statistics. 1991; 19:1.

31. Breiman L. Machine Learning. 2001; 45:5.

32. Witten, IH.; Frank, E. Morgan Kaufmann series in data management systems. 2. Morgan Kaufman; Amsterdam; Boston, MA: 2005. Data mining: practical machine learning tools and techniques; p. xxxip. 525

33. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Bioinformatics. 2004; 20:2479. [PubMed: 15073010]

34. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical Analysis Strategies for Association Studies Involving Rare Variants. Nature Reviews Genetics. 2010 In press.

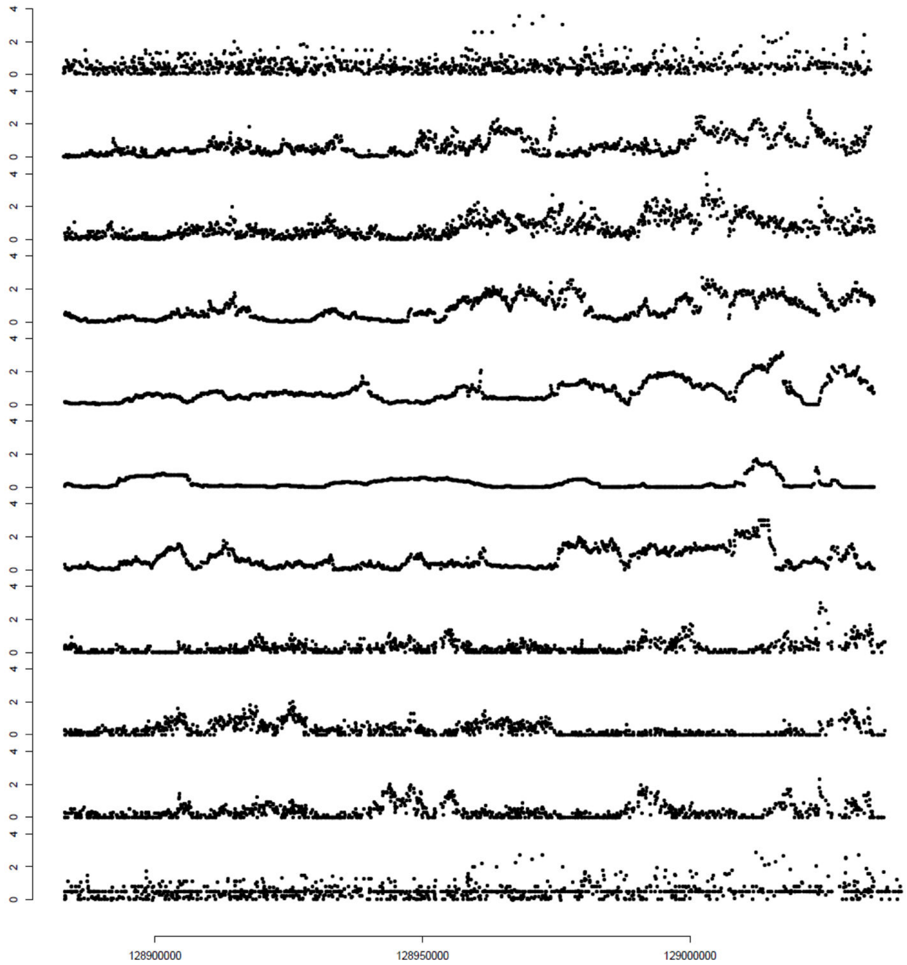

Figure 4.2.

Moving window analysis of the MGLL gene using 11 different methods. Note that the y axis provides the −log(p-value) for the association for all variants in the 5 kb window whose midpoint is given on the x axis.

Table 4.1

P-values for tests of the association of multiple variants within five functional genomic regions in the FAAH and MGLL genes.

	FAAH				
	NS	**H3K27**	**TFBS**	**FOX2**	**Amidase**
# of variants	**5**	**29**	**4**	**14**	**5**
Dispersion (Dis)	0.59	0.05	0.77	0.99	0.61
Diversity (Div)	0.43	0.42	0.81	0.33	0.46
MDMR Similarity (Sim)	0.19	0.21	0.05	0.14	0.41
Li & Leal (LL)	0.60	0.03	0.60	1.00	0.50
Subset Selection (SS)	1.00	0.01	0.60	0.75	0.60
Madsen & Browning (MB)	1.00	0.01	0.33	1.00	0.75
Logic Regression (LR)	0.23	0.18	0.39	0.22	0.48
Ridge Regresssion (RR)	0.35	0.09	0.06	0.33	0.54
PLINK Haplotype (Phap)	NA	0.92	NA	0.34	0.61
PLINK Set Analysis (Pset)	1.00	1.00	0.02	1.00	1.00

	MGLL				
	NS	**H3K27**	**TFBS**	**FOX2**	**Amidase**
# of variants	**9**	**100**	**11**	**3**	**0**
Dispersion (Dis)	0.28	0.99	0.02	0.72	NA
Diversity (Div)	0.77	0.65	0.73	0.64	NA
MDMR Similarity (Sim)	0.81	0.07	0.67	0.29	NA
Li & Leal (LL)	1.00	1.00	1.00	0.75	NA
Subset Selection (SS)	0.60	0.43	1.00	1.00	NA
Madsen & Browning (MB)	0.75	0.30	0.02	0.20	NA
Logic Regression (LR)	0.35	0.67	0.02	0.49	NA
Ridge Regresssion (RR)	0.71	0.50	0.01	0.61	NA
PLINK Haplotype (Phap)	NA	0.81	0.07	NA	NA
PLINK Set Analysis (Pset)	1.00	0.43	0.05	1.00	NA

Table 4.3

Spearman correlations between test statistics from the moving window analyses.

Method	Dis	Div	Sim	Ridge	Logic	LL	MB	SS
Dis		0.13	0.40	0.29	0.27	−0.01	−0.10	0.03
Div	0.13		−0.11	−0.09	−0.22	−0.01	−0.20	−0.11
Sim	0.40	−0.11		0.34	0.37	−0.06	−0.03	0.04
Ridge	0.29	−0.09	0.34		0.69	−0.08	−0.17	0.16
Logic	0.27	−0.22	0.37	0.69		−0.10	−0.23	0.00
LL	−0.01	−0.01	−0.06	−0.08	−0.10		0.25	0.49
MB	−0.10	−0.20	−0.03	−0.17	−0.23	0.25		0.21
SS	0.03	−0.11	0.04	0.16	0.00	0.49	0.21	

Table 4.4

Top 5 chosen genomic predictors of obesity for different regression analysis methods.

RR	L	GPS	SR	SPM-CART
166 (MGLL)	166 (MGLL)	166 (MGLL)	124 (FAAH)	1036 (MGLL)
677 (MGLL)	677 (MGLL)	677 (MGLL)	8 (FAAH)	1009 (MGLL)
581 (MGLL)	76 (MGLL)	76 (MGLL)	136 (FAAH)	H3K27 (MGLL)
76 (MGLL)	581 (MGLL)	428 (MGLL)	223 (FAAH)	1136 (MGLL)
90 (FAAH)	90 (FAAH)	90 (FAAH)	200 (FAAH)	H3K27 (FAAH)
adj. R2: 0.008	**adj. R2: 0.008**	**adj. R2: 0.011**	**adj. R2: <0**	**adj. R2: 0.066**
RSE: 10.56	**RSE: 10.56**	**RSE: 10.54**	**RSE: 10.62**	**RSE: 10.25**

SPM-TreeNet	SPM-MARS	SPM-RF	Weka CRL	Weka REPT
H3K27 (MGLL)	1036 (MGLL)	H3K27 (MGLL)	1058 (MGLL)	1036 (MGLL)
H3K27 (FAAH)	1009 (MGLL)	1036 (MGLL)	H3K27 (MGLL)	
1036 (MGLL)	654 (MGLL)	634 (MGLL)	56 (FAAH)	
1136 (MGLL)		H3K27 (FAAH)	210 (MGLL)	
1009 (MGLL)		632 (MGLL)	173 (FAAH)	
adj. R2: 0.066	**adj. R2: 0.076**	**adj. R2: 0.038**	**adj. R2: 0.033**	**adj. R2: 0.025**
RSE: 10.25	**RSE: 10.19**	**RSE: 10.4**	**RSE: 10.43**	**RSE: 10.47**