Published in final edited form as:

Genet Epidemiol. 2011 December ; 35(8): 880–886. doi:10.1002/gepi.20638.

On the Analysis of Sequence Data: Testing for Disease Susceptibility Loci using Patterns of Linkage Disequilibrium

Peter J Lipman¹, Wai-Ki Yip¹, Taofik AlChawa², Kerstin U Ludwig^{2,3}, Elisabeth Mangold², and Christoph Lange^{1,4}

¹Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

²Institute of Human Genetics, University of Bonn, Bonn, Germany

³Department of Genomics, Life and Brain Center, University of Bonn, Bonn, Germany

⁴Institute for Genomic Mathematics; German Center for Neurodegenerative Diseases, University of Bonn, Bonn, Germany

Abstract

Despite the numerous, successful applications of GWASs, there has been much difficulty in discovering DSLs. This is due to the fact that the GWAS approach is an indirect mapping technique, often identifying markers. For the identification of DSLs, which is required for the understanding of the genetic pathways for complex diseases, sequencing data that examines every genetic locus directly is necessary. Yet there is currently a lack of methodology targeted at the identification of the DSLs in sequencing data: existing methods localize the causal variant to a region, but not to a single variant and therefore do not allow one to identify unique loci that cause the phenotype association. Here, we have developed such a method to determine if there is evidence that an individual loci affects case-control status with sequencing data. This methodology differs from other rare variant approaches: rather than testing an entire region comprised of many loci for association with the phenotype, we can identify the individual genetic locus that causes the association between the phenotype and the genetic region. For each variant, the test determines if the pattern of LD across the other variants coincides with the pattern expected if that variant were a DSL. Power simulations show that the method successfully detects the causal variant, distinguishing it from other nearby variants (in high LD with the causal variant), and outperforms the standard tests. The efficiency of the method is especially apparent with small samples, which are currently realistic for studies due to sequence data costs. The practical relevance of the approach is illustrated by an application to a sequence dataset for nonsyndromic cleft lip with or without cleft palate. The proposed method implicated one variant ($p=0.002$, .062 after Bonferroni correction), which was not found by standard analyses. Code for implementation is available.

Introduction

Since 2004, major improvements in genotyping technology have lead to the large scale production of inexpensive SNP-chips with genome-wide coverage. The technological development has reduced the costs and the laboratory efforts for genome-wide association studies (GWAS) so much that they have become a standard research tool for disease gene

mapping. Numerous, successful applications of GWAS to complex diseases and phenotypes have confirmed their effectiveness. GWAS have identified many novel genetic associations with complex phenotypes that can be reliably replicated in independent studies [Manolio 2010]. Although the successes of GWAS are undoubtedly a major step forward in the disease mapping process, GWAS have not led to the discovery of many disease susceptibility loci (DSLs). This shortcoming is attributable to the fact that the GWAS approach is an indirect mapping technique, i.e. genetic loci that are in linkage disequilibrium (LD) with the DSL are detected, but not the actual DSLs. For the identification of DSLs - which is required for the understanding of the genetic pathways for complex diseases and phenotypes-sequencing data is necessary. With sequencing data at hand, every genetic locus can be examined directly. Currently, the research focus is on the development of high-throughput sequencing technology. Although sequencing is still expensive and not a standard research tool, the costs of sequencing will decrease substantially over the next few years, introducing this technology into mainstream genetic research.

To translate the wealth of information in sequencing data successfully into DSL discovery, new statistical approaches are needed. The majority of genetic loci that are recorded by sequencing are rare variants, i.e. loci with minor allele frequencies (MAF's) less than 1%. Empirical and theoretical evidence [Nejentsev et al 2009] [Fearnhead et al. 2004] [Cohen et al. 2006] [Ji et al. 2008] [Ingason et al. 2011] [Stefansson et al. 2008] [Walsh et al. 2008] [Weiss et al. 2008] [Helbig et al. 2009] [Hill et al. 1994] [Pritchard et al. 2002] [Pritchard 2001] [Kryukov et al. 2007] [Adzhubei et al. 2010] suggests that associations signals that are detected by GWAS can be caused by multiple, rare variants that are in the vicinity of the SNPs that were identified by the GWAS. Although several approaches have already been developed for designs of unrelated individuals, e.g. case/control studies [Bansal et al. 2010] [Li et al. 2008] [Madsen et al. 2009] [Price et al. 2010] [Yi et al. 2011], there is a lack of methodology that is targeted at the identification of the DSLs in sequencing data. Existing methodology localizes the DSL to a region, but not to a single variant and therefore does not allow one to identify unique loci that cause the phenotype association. The methodology developed here is different from all other rare variant approaches, as we test for phenotype association with individual loci, rather than over a region comprised of many loci. The methodology is most appropriately used to detect DSLs after identifying candidate genes/regions. It is particularly useful in next generation sequencing, where many rare variants will be genotyped.

Nevertheless, the major goal of sequence-analysis is the identification of the DSLs. The significance of single-locus association tests is defined by the genetic effect size and the allele frequency. Since non-DSLs that are in LD with the true DSL can have higher allele frequencies than the DSL, but have smaller observed genetic effect sizes, the significance of the original (GWAS) non-DSL test cannot be used to identify DSLs. In order to distinguish the true DSLs from those SNPs that are correlated (i.e. in LD) with the DSLs, statistical approaches that assess differences in LD-pattern across multiple loci between subjects are used.

Methods

To identify the locations of the DSLs with high-throughput sequencing data, we propose a method that examines the differences in LD patterns between cases and controls, and between loci. We assume that we have sequence data from a case-control study on $i = 1, \dots, n$ variants. With this study design, we can estimate each variant's LD pattern across the other variants within each disease status group.

There are many measurements of LD, notably D , Lewontin's D' (or D'), the correlation coefficient R , and R^2 [Devlin et al. 1995]. The proposed method holds for any measurement of LD, but the choice of LD measure might have an influence on the power of the approach. We will examine the performance of the approach for different LD-measurements by simulation studies. For the following derivation, we assume that one of the LD-measurements has been selected and is used throughout the approach. Let m_{ij}^{ca} estimate the LD in the cases between the i^{th} and j^{th} variant, $i \neq j$. Let m_{ij}^{co} estimate the LD in the controls between the i^{th} and j^{th} variant, $i \neq j$. We define $m_{ij} = m_{ij}^{ca} - m_{ij}^{co}$, the difference between the cases and the controls. For each variant i , we will investigate if the LD across variants j follows a pattern that would suggest that variant i is a DSL.

First, we must determine how a DSL will be correlated (i.e. in LD) with other nearby variants. As we will see, the LD-pattern for a DSL depends upon whether the DSL and the other variants (for which we calculate LD) are rare or common. We first note that the minor allele for a DSL will be more often present among the cases, as the DSL is the causal variant and we assume the DSL has a detrimental effect. If the DSL is common, we expect it to be in higher LD with other common variants in the cases compared to the controls. This statement holds because minor allele of the DSL and of the other variant have the opportunity to "travel" together among the cases (since they are both common), but will not travel together among the controls (since the DSL, by definition, will not be as often present in the controls). The fact that common DSLs will be more highly correlated with other common variants in the cases than in the controls can also be argued by contradiction. We also expect a common DSL to be in lower LD with other rare variants in the cases compared to the controls for similar reasons: since the DSL is common but the other variant is rare, the two variants do not have the opportunity to travel together in the cases (since the minor allele of the DSL is more often in the cases by definition, and a more common variant will be less highly correlated with a rare variant), but they do have the opportunity to travel together among the controls, so LD will be higher in the controls. For the same reason, we also expect rare DSLs to be in lower LD with other common variants in the cases compared to the controls. Lastly, for a rare DSL and other rare variants, we expect the LD among cases to be different from the LD among controls, although the directionality is unknown. Our initial simulations and preliminary data (see below) show these patterns to hold under realistic settings. Statistical approaches for the localization of DSLs can take advantage of this pattern. Our observations about different LD patterns allow us to formulate a powerful test statistic to capture to what extent each variant follows the LD pattern of a DSL.

Based on these considerations, when variant i is common and a DSL, we expect $m_{ij} < 0$ if variant j is rare and $m_{ij} > 0$ otherwise. When variant i is rare and a DSL, we expect $m_{ij} < 0$ if variant j is common. The last scenario possible, if variant i is rare and is a DSL and variant j is rare as well, we expect differential patterns of LD between cases and controls, but the direction is unknown, i.e. we expect $m_{ij} \neq 0$. We define δ to be a particular MAF chosen to categorize variants as rare or common. These statements are summarized in Table 1 below.

We use these expected patterns of LD for a DSL to formulate a viable test statistic. For each variant i , the test statistic adds across other variants j , switching the sign of m_{ij} so that it is expected to be positive or large according to Table 1. The test statistic is thus created so that large values provide evidence that variant i is a DSL. The test statistic, T_i , for each variant i , is:

$$\begin{aligned} \text{If } \text{MAF}_i > \delta, \quad T_i &= \sum_{j \neq i} (-1)^{I\{\text{MAF}_j < \delta\}} m_{ij} \\ \text{If } \text{MAF}_i < \delta, \quad T_i &= \sum_{j \neq i} (|m_{ij}|)^{I\{\text{MAF}_j < \delta\}} (-m_{ij})^{I\{\text{MAF}_j \geq \delta\}} \end{aligned}$$

where $I\{\}$ is the indicator function.

In order to test if T_i is significantly larger than we would expect under the null hypothesis that variant i is not a DSL, we must determine the null distribution of the test statistic for each variant. This can be calculated by permuting the link between phenotype (disease status) and genotype. This leaves the LD pattern in tact for each subject but changes each person's disease status, and avoids the assumption that the sequence data follows any particular model. For each permutation, the test statistics are recalculated. Based on a sufficient number of permutations, the empirical distribution of the proposed test statistic under the null hypothesis can be obtained. The observed test statistics can then be compared to the null distribution. By construction of the test statistic, we perform a one-sided test for statistical significance (T_i large are significant). We can use a Bonferroni correction to account for the multiple testing of the n variants or permutation-based thresholds that take the correlation between the test statistics T_i into account.

Preliminary Simulation Studies

While the method proposed here holds for any measurement of LD, we focus in this simulation study on using R^2 , the square of the correlation coefficient. To test the power of the method under realistic conditions, sequence data was simulated using FREGENE [Chadeau-Hyam et al. 2008]. This C^{++} code allows for a wide range of scenarios for selection, recombination, migration, and population structure, as the simulations run forward in time. Included in the program are ready-to-use simulated test datasets. Here, we use the final generation of a neutral panmictic population, which is modeled with 10.5K individuals over 200,000 generations and simulated over 20Mb genomes. Parameters for the sequence simulations are detailed in [Chadeau-Hyam et al. 2008]. We simulated case-control status with disease prevalence of 10% based on sequence data with 100 variants. We chose to have this phenotype independently caused by five DSLs. The disease model assumes that the odds of disease are additive for each risk allele on the log scale, as in logistic regression. In

the first set of simulations, we assume the disease is caused by somewhat rare variants with more moderate risk. In the second set of simulations, we assume the disease is caused by very rare variants with larger effect sizes. The minor allele frequencies and odds ratios for each DSL in these two simulation scenarios are described in Table 2 and Table 3, respectively. With these parameters, we created a 15,000-person population, of which 5,000 were cases and 10,000 were controls.

For power simulations, first 500 cases and 500 controls were sampled from the large population over 1,000 simulations. For each simulation, the test statistic was computed for all 100 variants, and the null distribution was simulated using 10,000 permutations. For the first set of simulations, after Bonferroni-correction for 100 multiple tests at the $\alpha = 0.05$ level, we achieved, on average, 89.96% power for detecting the DSLs when the MAF cutoff (δ) was set to 0.001. We repeated the simulations with the MAF cutoff (δ) set to 0.005 then 0.01 and achieved, on average, 90.76% then 91.60% power, respectively. Power increased for the second set of simulations for (δ) set to 0.001 and 0.005. These results are listed in Table 4. We repeated the simulations with the same parameters described above, sampling 2,500 cases and 2,500 controls. After Bonferroni-correction for 100 multiple tests at the $\alpha = 0.05$ level, we achieved, on average, 100% power for detecting the DSLs when the MAF cutoff (δ) was set to 0.001, 0.005, or 0.01. For the second set of simulations, power somewhat declined for $\delta = 0.01$. Again, results are summarized below in Table 4.

In order to compare our new method to current association tests for DSL detection, we computed the Armitage Trend test as well as the Fisher Exact test, since the Armitage Trend test may not be appropriate when the cell counts are small. We note that the only methods that are directly comparable to our approach are the single locus tests, so we only compare our method to other single locus tests. Over 10,000 simulations, for the 500-case/500-control study, average power for the Armitage Trend test was 87.61% after Bonferroni-correction and average power for the Fisher Exact test was 85.51% after Bonferroni-correction in the first set of simulations. For the 2,500-case/2,500-control scenario, average power for both the Armitage Trend test and Fisher Exact test were 100.00% after Bonferroni-correction. For the second set of simulations, power somewhat declined for the 500-case/500-control scenario. These results are displayed in Table 4. Overall, we see that the standard tests have lower power than the proposed LD method, and that the power of the standard tests declines as the DSLs become more rare, while the power of the proposed LD method increases.

For both baseline comparison tests, it is important to note that, for the 95 variants that were not DSLs, they were identified in the first set of simulations as statistically significant 7.04% (18.31%) and 5.59% (17.48%) of the time for the 500-case/500-control scenario (2,500-case/2,500-control) for the Armitage Trend test and Fisher Exact test, respectively. These tests could have therefore led to the wrong conclusion about the location of the true DSL, as the type-1 error/ α -level is inflated, especially for larger samples sizes. This feature of the tests is likely due to the strong LD between rare variants and the true DSLs, and we label this quantity as the "detection percentage" (of the non-DSLs). In contrast, for our test statistic, variants that were in LD with the true DSL were identified as statistically significant in the first set of simulations only 2.39% (2.63%, 2.71%) of the time when the MAF cutoff (δ) was set to 0.001 (0.005, 0.01) for the 500-case/500-control scenario and

3.91% (4.26%, 4.33%) for the 2,500-case/2,500-control scenario. The results are similar in the second set of simulations as well, and are also reported in Table 4. We do note here that the power of the method, as with any statistical test, will depend upon the integrity of the data. Genotyping errors will decrease power for our method, as for any statistical test. We also note here that, to be anti-conservative, we used a no-selection model in our simulation studies. In the presence of selection, the LD-patterns will be stronger, and our method will perform better in such scenarios. Single-locus tests do not compare the LD between loci. Consequently, their performance for DSL detection will be not affected by the changes in LD-patterns between loci. Furthermore, we feel that for most complex diseases the absence of selection is reasonable, especially for rare variants.

We also graphically show in Figure 1 how the different methods perform for each of the 100 SNPs individually for the 500-case/500-control scenario, comparing the Armitage Trend test (standard method) to the LD method with $\delta = .001$. Graphical results for the Fisher Exact test are similar to those shown for the Armitage Trend test, while graphical results for the LD method with different δ values are similar to those shown in Figure 1 in all simulations.

Results - Data Application

We now show the practical relevance of the approach with an application to a case-control sequence dataset. With a prevalence of 1/700 to 1/1000, cleft lip with or without cleft palate (CL/P) is one of the most common congenital malformations in European populations. It may occur as part of a complex malformation syndrome or as an isolated anomaly (nonsyndromic CL/P; NSCL/P). Nonsyndromic cases, which represent approximately 70% of all CL/P cases, are considered to have a multifactorial etiology which involves both genetic and environmental factors [Mossey et al. 2009].

The resequencing study involved 96 NSCL/P patients and 96 controls of Central European origin (Germany and neighboring countries). Among the index patients, 29 have a positive family history (two or more NSCL/P affected individuals in at least two consecutive generations). All patients had been examined by one of two medical geneticists to exclude any underlying syndrome. A detailed questionnaire to identify any possible prenatal contributory factors had been completed. No maternal ingestion of known teratogenic medications or toxins had been reported. Controls (volunteer blood donors) had not been screened for cleft status, but this should not result in any appreciable reduction in power since the prevalence of orofacial clefting in the general population is very low [Moskvina et al. 2005]. EDTA anticoagulated peripheral venous blood samples were collected from all individuals. Lymphocyte DNA was isolated using a standard salting out procedure. The study was approved by the ethics committees of the Universities of Bonn and Gottingen.

Evidence for the candidate gene sequenced in this study comes from a genome wide association study that was recently published [Mangold et al. 2010]. In that study, using 399 NSCL/P cases and 1318 controls loci, three loci with suggestive evidence of association at the genome-wide significance level were identified. The candidate gene resequenced in this study is located at one of these suggestive loci (15q13.3; manuscript in preparation). Sequencing comprised the entire coding region and untranslated regions and was performed

using on an Applied Biosystems 3130XL Genetic Analyzer (Applied Biosystems, California, USA). Primer sequences and PCR conditions are available upon request after publication of the original data.

We first applied the typical sequencing data analysis approach of collapsing over the region [Li et al. 2008] to determine if there was any evidence for association of the entire genetic region and case-control status. This test was not significant, illustrating that the approach is not suitable for DSL detection. We then applied our method to the 27 variants that were identified, mostly rare, with MAF's described in Table 5 below. Standard analysis using the Armitage Trend Test reveals no significant findings after Bonferroni correction, as shown in Table 6. In order to apply the LD method, haplotypes were inferred using PHASE 2.1.1 [Stephens et al. 2001] [Stephens et al. 2003] with 92.2% certainty. Cases and controls were inferred together to conserve the type-1 error rate, as discussed by Balding [2006]. The LD method, with $\delta=0.001$, finds one suggestively significant p-value at SNP 15 ($p=0.06$ after Bonferroni-correction or FDR-correction). Results for the top 5 SNPs are reported in Table 6. The genotype distribution for SNP 15 is displayed in Table 7. For this SNP, the Armitage Trend Test is significant before correcting for multiple testing, with an estimated odds ratio of 2.76, 95% CI = (1.055,7.24). This application demonstrates the ability of our new methodology to identify causal genetic loci and illustrates the limitation of the standard analysis in this context. We identify that SNP 15 is driving the signal identified by the original GWAS ($p=0.06$) and likely the causal variant. We note that no other methods could have implicated this variant, as the standard methods lack power for sequencing data, and other existing methodology only tests for associations among regions, not single variants. We recommend further functional assessment of this variant.

Discussion

Numerous, successful applications of GWAS's have identified novel genetic associations with complex phenotypes that can be reliably replicated in independent studies [Manolio 2010], but have not led to the discovery of many DSLs. This shortcoming is attributable to the fact that the GWAS approach is an indirect mapping technique, detecting loci that are in LD with the DSL, but not the actual DSLs. For the identification of DSLs - which is required for the understanding of the genetic pathways for complex diseases and phenotypes - sequencing data, which examines every genetic locus directly, is necessary.

The costs of sequencing will decrease substantially over the next few years, introducing this technology into mainstream genetic research. Yet there is currently a lack of methodology that is targeted at the identification of the DSLs in sequencing data, as methodology localizes the DSL to a region, but not to a single variant. Here, we have developed such a method to determine if there is evidence that an individual variant affects case/control status with sequence data. For each variant, the test determines if the pattern of LD across the other variants coincides with the pattern expected if that variant were a DSL. The test is computationally straightforward, using a permutation strategy. We are currently working on extending the methodology to quantitative traits.

We have demonstrated that the method achieves adequate power in realistic scenarios and outperforms the standard tests (Armitage Trend and Fisher Exact), especially for very rare variants. An important feature of the test is its ability to distinguish between the true DSL and other variants that are in LD with the DSL, maintaining the type-1 error rate even for those variants that are in high LD with the DSL. Although the test analyzes each marker individually, it uses information across all variants to calculate LD patterns, thus adding efficiency. Compared to other methods, the LD method is superior when sample sizes are small. This is when the gain in efficiency is most important. Currently, realistic sequence datasets involve smaller samples. Even for large sample sizes, simulations show that the LD method maintains the same power as the standard analysis techniques, while still having a lower detection percentage of SNPs that are not the DSLs, but are instead closely linked (i.e. markers).

The method was applied to a case/control dataset for a gene previously implicated in a GWAS for nonsyndromic cleft lip with or without cleft palate. The standard sequencing data analysis technique by Li [2008] does not find significant association between the entire genetic region and case-control status, suggesting the method is not ideal for DSL detection. The proposed LD method, with $\alpha=0.001$, finds one suggestively significant p-value ($p=0.002$ without Bonferroni correction and $p=0.06$ after Bonferroni-correction). For the same SNP, the Armitage Trend Test has a p-value of $p=0.03$ before Bonferroni-correction and $p=0.83$ after Bonferroni correction. This application illustrates the practical advantages of the proposed method. We note that no other methods could have implicated this variant, as the standard methods lack power for sequencing data, and other existing methodology only tests for associations among regions, not single variants. Code for implementation using the open software R or a standalone C/C++ program (on Linux64 and Win32 platforms) is currently available and may be found at <http://people.hsph.harvard.edu/~plipman>. The implementation will also be part of a larger population based statistical software package, *N P BAT*, which will be available by the end of summer 2011.

Acknowledgments

We would like to thank Dr. Clive J. Hoggart for his help in using FRE-GENE/SAMPLE.

For the NSCL/P study, we thank all affected individuals and their families for participation in this study, and all collaborating clinical partners for their contribution of DNA samples. The control DNA samples were kindly provided by Prof. Bernd Potzsch (Institute of Experimental Hematology and Transfusion Medicine, University of Bonn). Nadine Kluck is acknowledged for technical assistance. The NSCL/P study was supported by the Deutsche Forschungsgemeinschaft (FOR 423 and individual grants MA 2546/3-1, KR 1912/7-1, NO 246/6-1 and WI 1555/5-1). Taofik AlChawa is supported by a grant from the Ministry of Higher Education, Syrian Arab Republic.

Funding: This work was supported by NIH RO1 MH087590 and RO1 MH081862.

References

- [Adzhubei et al 2010]. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010 Apr; 7(4):248–9. [PubMed: 20354512]
- [Balding 2006]. Balding DJ. A Tutorial on Statistical Methods for Population Association Studies. *Nature Genetics*. 2006; 7:783.

- [Bansal et al 2010]. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet.* 2010 Nov; 11(11):773–85. [PubMed: 20940738]
- [Chadeau-Hyam et al 2008]. Chadeau-Hyam M, Hoggart CJ, Reilly PF, Whittaker JC, De Iorio M, Balding DJ, Fregene. Simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics.* 2008; 9:364.
- [Cohen et al 2006]. Coffman FD, He M, Diaz ML, Cohen S. Multiple initiation sites within the human ribosomal RNA gene. *Cell Cycle.* 2006 Jun; 5(11):1223–33. [PubMed: 16721058]
- [Devlin et al 1995]. Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics.* 1995; 29:311–322. [PubMed: 8666377]
- [Fearnhead et al 2004]. Fearnhead NS, Wilding JL, Winney B, Tonks S, Bartlett S, Bicknell DC, Tomlinson IP, Mortensen NJ, Bodmer WF. Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc Natl Acad Sci U S A.* 2004 Nov 9; 101(45):15992–7. [PubMed: 15520370]
- [Helbig et al 2009]. Helbig I, Mefford HC, Sharp AJ, Guipponi M, Fichera M, Franke A, Muhle H, de Kovel C, Baker C, von Spiczak S, Kron KL, Steinich I, Kleefuss-Lie AA, Leu C, Gaus V, Schmitz B, Klein KM, Reif PS, Rosenow F, Weber Y, Lerche H, Zimprich F, Urak L, Fuchs K, Feucht M, Genton P, Thomas P, Visscher F, de Haan GJ, Mller RS, Hjalgrim H, Luciano D, Wittig M, Nothnagel M, Elger CE, Nrnberg P, Romano C, Malafosse A, Koeleman BP, Lindhout D, Stephani U, Schreiber S, Eichler EE, Sander T. 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nat Genet.* 2009 Feb; 41(2):160–2. [PubMed: 19136953]
- [Hill et al 1994]. Hill WG, Weir BS. Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am J Hum Genet.* 1994 Apr; 54(4):705–714. [PubMed: 8128969]
- [Ingason et al 2011]. Ingason A, Rujescu D, Cichon S, Sigurdsson E, Sigmundsson T, Pietilinen OP, Buizer-Voskamp JE, Strengman E, Francks C, Muglia P, Gylfason A, Gustafsson O, Olason PI, Steinberg S, Hansen T, Jakobsen KD, Rasmussen HB, Giegling I, Mller HJ, Hartmann A, Crombie C, Fraser G, Walker N, Lonnqvist J, Suvisaari J, Tuulio-Henriksson A, Bramon E, Kiemeny LA, Franke B, Murray R, Vassos E, Touloupoulou T, Mhleisen TW, Tosato S, Ruggeri M, Djurovic S, Andreassen OA, Zhang Z, Werge T, Ophoff RA, Rietschel M, Nthen MM, Petursson H, Stefansson H, Peltonen L, Collier D, Stefansson K, St Clair DM. GROUP Investigators. Copy number variations of chromosome 16p13.1 region associated with schizophrenia. *Mol Psychiatry.* 2011 Jan; 16(1):17–25. [PubMed: 19786961]
- [Ji et al 2008]. Ji W, Foo JN, O’Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet.* 2008 May; 40(5):592–9. [PubMed: 18391953]
- [Kryukov et al 2007]. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare mis-sense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet.* 2007 Apr; 80(4):727–39. [PubMed: 17357078]
- [Li et al 2008]. Li B, Leal SM. Novel Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *American Journal of Human Genetics.* 2008; 83:311–21. [PubMed: 18691683]
- [Madsen et al 2009]. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009 Feb; 5(2)
- [Mangold et al 2010]. Mangold E, Ludwig KU, Birnbaum S, Baluado C, Ferrian M, Herms S, Reutter H, de Assis NA, Chawa TA, Mattheisen M, Steffens M, Barth S, Kluck N, Paul A, Becker J, Lauster C, Schmidt G, Braumann B, Scheer M, Reich RH, Hemprich A, Ptzsch S, Blaumeiser B, Moebus S, Krawczak M, Schreiber S, Meitinger T, Wichmann HE, Steegers-Theunissen RP, Kramer FJ, Cichon S, Propping P, Wienker TF, Knapp M, Rubini M, Mossey PA, Hoffmann P, Nthen MM. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet.* 2010; 42:24–26. [PubMed: 20023658]
- [Manolio 2010]. Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med.* 2010 Jul 8; 363(2):166–76. [PubMed: 20647212]
- [Moskvina et al 2005]. Moskvina V, Holmans P, Schmidt KM, Craddock N. Design of case-controls studies with unscreened controls. *Ann Hum Genet.* 2005; 69:566–576.

- [Mossey et al 2009]. Mossey PA, Little J, Munger RG, Dixon MJ, Shaw WC. Cleft lip and palate. *Lancet*. 2009; 374:1773–1785. [PubMed: 19747722]
- [Nejentsev et al 2009]. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*. 2009 Apr 17; 324(5925):387–9. [PubMed: 19264985]
- [Price et al 2010]. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet*. 2010 Jun 11; 86(6):832–8. [PubMed: 20471002]
- [Pritchard 2001]. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*. 2001 Jul; 69(1):124–37. [PubMed: 11404818]
- [Pritchard et al 2002]. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet*. 2002 Oct 1; 11(20):2417–23. [PubMed: 12351577]
- [Stefansson et al 2008]. Stefansson H, Rujescu D, Cichon S, Pietilinen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Miller HJ, Hartmann A, Shianna KV, Ge D, Need AC, Crombie C, Fraser G, Walker N, Lonnqvist J, Suvisaari J, Tuulio-Henriksson A, Paunio T, Touloupoulou T, Bramon E, Di Forti M, Murray R, Ruggeri M, Vassos E, Tosato S, Walshe M, Li T, Vasilescu C, Mhleisen TW, Wang AG, Ullum H, Djurovic S, Melle I, Olesen J, Kiemeny LA, Franke B, Sabatti C, Freimer NB, Gulcher JR, Thorsteinsdottir U, Kong A, Andreassen OA, Ophoff RA, Georgi A, Rietschel M, Werge T, Petursson H, Goldstein DB, Nthen MM, Peltonen L, Collier DA, St Clair D, Stefansson K. GROUP. Large recurrent microdeletions associated with schizophrenia. *Nature*. 2008 Sep 11; 455(7210):232–6. [PubMed: 18668039]
- [Stephens et al 2001]. Stephens M, et al. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*. 2001; 68:978–989. [PubMed: 11254454]
- [Stephens et al 2003]. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction. *American Journal of Human Genetics*. 2003; 73:1162–1169. [PubMed: 14574645]
- [Walsh et al 2008]. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Rocanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*. 2008 Apr 25; 320(5875):539–43. [PubMed: 18369103]
- [Weiss et al 2008]. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, Platt OS, Ruderfer DM, Walsh CA, Altshuler D, Chakravarti A, Tanzi RE, Stefansson K, Santangelo SL, Gusella JF, Sklar P, Wu BL, Daly MJ. Autism Consortium. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med*. 2008 Feb 14; 358(7):667–75. [PubMed: 18184952]
- [Yi et al 2011]. Yi N, Zhi D. Bayesian analysis of rare variants in genetic association studies. *Genet Epidemiol*. 2011 Jan; 35(1):57–69. [PubMed: 21181897]

Figure 1.

Detection probabilities for 500 cases/500 controls in simulation 1 by each of the 100 variants. Here, the black bars represent the proper rejection of the null hypothesis for the DSLs (i.e. power), while the white bars indicate the improper rejection of the null hypothesis for non-DSLs, likely due to strong LD among the SNPs (i.e. type-1 error or identifying markers rather than DSLs). On the left are the results for each variant using the standard Armitage trend test. On the right are the results for the new proposed method, with $\delta=0.001$. We see that in the new proposed method, the black bars become elevated (i.e. power is increased), while the white bars lower (i.e. fewer type-1 errors, or less identification of markers rather than DSLs). Graphical results for the Fisher Exact test are similar to those shown for the Armitage Trend test (left), while graphic results for the LD method with different δ values are similar to those shown for the LD method (right) in both simulation scenarios.

Table 1

Expected patterns of LD between DSLs and other variants, where δ is a MAF that dichotomizes variants as rare or common and m_{ij} is the difference in LD between the cases and the controls for variants i and j

	DSL i common $\text{MAF}_i \geq \delta$	DSL i rare $\text{MAF}_i < \delta$
variant j common $\text{MAF}_j \geq \delta$	$m_{ij} > 0$	$m_{ij} < 0$
variant j rare $\text{MAF}_j < \delta$	$m_{ij} < 0$	$m_{ij} \geq 0$

Table 2

DSL attributes for 1st simulation

SNP	87	30	74	41	29
MAF	.290	.015	.006	.006	.009
OR	1.2	3	4	5	6

Table 3

DSL attributes for 2nd simulation

SNP	30	46	41	38	79
MAF	.015	.011	.006	.003	.002
OR	5	6.25	7.5	8.25	10

Table 4

Power (P-) and detection percentage (DP-) of non-DSLs by method; P-1 and DP-1 for first simulation, P-2 and DP-2 for second simulation

Sample Size	Test	P-1	P-2	DP-1	DP-2
500/500	Armitage Trend	87.61%	84.52%	7.04%	8.00%
500/500	Fisher Exact	85.51%	84.54%	5.59%	7.93%
500/500	LD, $\delta=0.001$	89.96%	93.64%	2.39%	6.00%
500/500	LD, $\delta=0.005$	90.76%	91.00%	2.63%	6.00%
500/500	LD, $\delta=0.01$	91.60%	72.82%	2.71%	5.11%
2500/2500	Armitage Trend	100.00%	100.00%	18.31%	11.00%
2500/2500	Fisher Exact	100.00%	100.00%	17.48%	10.54%
2500/2500	LD, $\delta=0.001$	100.00%	100.00%	3.91%	8.74%
2500/2500	LD, $\delta=0.005$	100.00%	100.00%	4.26%	8.62%
2500/2500	LD, $\delta=0.01$	100.00%	77.90%	4.33%	8.24%

Table 5

Summary of MAF's for the 27 variants

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0026	0.0026	0.0287	0.1296	0.2461	0.4375

Table 6

P-values of SNPs on disease status by different methods; top 5 of 27 reported

SNP	LD	LD Bonf-corrected	Armitage	Armitage Bonf-corrected
15	0.002	0.062	0.031	0.836
17	0.014	0.370	0.163	1.000
19	0.006	0.175	0.048	1.000
25	0.068	1.000	0.067	1.000
27	0.024	0.659	0.191	1.000

Table 7

Genotype distribution for SNP 15

Disease Status	CC	CT	TT
Control	90	6	0
Case	81	14	1