Published in final edited form as:

Methods Mol Biol. 2012 ; 850: 453–464. doi:10.1007/978-1-61779-555-8_24.

Detecting rare variants

Tao Feng* and Xiaofeng Zhu

Summary

The limitations of genome-wide association (GWA) studies that are based on the common disease common variants (CDCV) hypothesis have motivated geneticists to test the hypothesis that rare variants contribute to the variation of common diseases, i.e., common disease rare variants (CDRV). The newly developed high-throughput sequencing technologies have made the studies of rare variants practicable. Statistical approaches to test associations between a phenotype and rare variants are quickly developing. The central idea of these methods is to test a set of rare variants in a defined region or regions by collapsing or aggregating rare variants, thereby improving the statistical power. In this chapter, we introduce these methods as well as their applications in practice.

Keywords

GWA; Common disease common variants; Common disease rare variants; SNPs; Haplotype; Collapsing; Aggregation

1. Introduction

1.1 Recent Successes, Limitations and Challenges of Genome-Wide Association Studies

The underlying biology of common diseases is unclear. It has been suggested that both genetic and environmental factors contribute to the variation in risk of complex diseases. Studying the genetic risk factors for a common disease will be helpful for understanding the etiology of the disease and, therefore, to prevent the disease eventually. The completion of the human genome sequence (1, 2), and the progress in SNP genotyping technologies have facilitated the creation of dense SNP databases such as the International HapMap project (3, 4). Many genome-wide association (GWA) studies have been completed and several genetic variants have been identified as being associated with complex diseases. A hypothesis underlying GWA studies is that causal variants are common in a population (5, 6). Under this assumption, variants contributing to the variation in risk of common disease susceptibility can be detected by testing tagging SNPs across the genome. A well-known whole genome association study, the Wellcome Trust Case Control Consortium (WTCCC) study (7), was based on the CDCV assumption. By comparing the allele frequencies between cases and controls in the WTCCC study, 24 independent association signals were identified. These signals were 1 for Bipolar Disorder, 1 for Coronary Artery Disease, 9 for

*Corresponding author: Department of Epidemiology and Biostatistics, Case Western Reserve University School of Medicine, Cleveland, Ohio 44106, USA, Phone: 216-368-5659, tfeng@darwin.case.edu.

Crohn's disease, 3 for Rheumatoid Arthritis, 7 for Type I diabetes and 3 for Type II diabetes (7).

Despite the remarkable successes in identifying genetic variants that influence common diseases, many challenges still lie ahead. The majority of genetic variants contributing to disease susceptibility are yet to be discovered. Some common diseases, such as hypertension, are known to have strong genetic components. Yet uncovering their underlying putative genetic variants using GWA studies has proven to be difficult. The common variants that have been identified to date can only explain a small proportion of the variation in disease phenotype. For example, height is known to be a heritable trait with estimated heritability around 0.8, which is about 80% of the inter-individual variation attributable to genetic factors. Although the three GWA studies in 2008 (8, 9, 10) have found 40 previously unknown variants, each locus can only explain a small proportion of the phenotypic variance (0.3% - 0.5%). The failure of GWA studies to identify the majority of susceptibility variants can be attributed to various aspects, such as population stratification, small effect sizes of genetic variants, copy number variations and rare variants.

1.2 Multiple Rare Variants Hypothesis and Recent Evidence

Some recent studies persuade us to believe that multiple rare variants may contribute to disease risk in the human population. Recent findings, such as the association of *BRCA1* and *BRCA2* to breast cancer (11), indicates that multiple rare variants within the same gene can contribute to largely monogenic disorders. Most of the common variants identified in GWA studies have modest effective sizes with odds ratios between 1.2 and 1.5 (12). It has been suggested that rare variants may have larger effective sizes than the common variants (13). However, the analysis of rare variants can be challenging. In this Chapter, we describe some recently developed methods for analyzing rare genetic variants.

There are several approaches to test the association of rare variants and a disease outcome or trait. One approach is to test each variant individually using a standard contingency table or regression method. However, this approach has low statistical power to detect an association between the rare variants and the trait owing to their small allele frequencies (14, 15, 16). To overcome this issue, one viable strategy is to “collapse or aggregate” sets of rare variants into a single group and test their collective frequency difference between cases and controls (15, 17).

We first introduce some notation and explain different strategies for collapsing the rare variants. We provide specific details for using a two-stage haplotype based method for evaluating rare variants. Next, in the Methods section, we explain how to conduct the two-stage haplotype based analysis using the R programming language (<http://cran.r-project.org>). We focus on case-control association studies, where cases are individuals who have the disease of interest, and controls are disease-free individuals. We assume that the cases and controls are independent. We also refer to the cases as affected individuals, and the controls as unaffected individuals.

Notation—Assume that within a region there are M SNPs that can independently cause disease susceptibility. The term “region” refers to the unit in which the variants will be

collectively analyzed. It can be defined by using a gene or pathway, or a genomic block based on some criteria. Each SNP has two alleles, denoted A_i and a_i , $i = 1, 2, \dots, M$, in which A_i always refers to the rare high-risk allele and has an allele frequency p_i . Furthermore, let G_k , $k = 0, 1, 2$ denote the genotypes aa , Aa , and AA , respectively.

1.3 Collapsing Method

Our goal is to test whether the presence of rare variants in a region is associated with disease. Define an indicator variable X for the j th case individual as,

$$X_j = \begin{cases} 1 & \text{rare variant(s) present} \\ 0 & \text{otherwise.} \end{cases}$$

Y_j is defined similarly for control individuals. Due to the rarity of variants, the probability of carrying more than one variant for an individual is low (see Note 1), and the method collapses genotypes across all variants, such that an individual is coded as 1 if a rare allele is present at any of the variant sites in the region M and as 0 otherwise. The detection of an association of multiple rare variants is transformed into a test of whether the proportions of individuals with rare variants in cases and controls differ. Any single SNP association test that is applied in GWA studies can be applied here, such as a chi-squared test for a contingency table or a regression analysis.

1.4 Combined Multivariate and Collapsing (CMC)

Li and Leal (15) considered an extension of the collapsing method, which they termed the combined multivariate and collapsing (CMC) method, to take advantage of both the multiple marker tests and the collapsing method. For a predefined region, they first divide the markers (for example, SNPs) within the region into groups according certain criteria (for example, allele frequencies) and then collapse the rare variants within each group using the **collapsing method** described above. To analyze the groups of collapsed rare variants, a multivariate test such as Hotelling's T^2 test is applied.

1.5 Weighted Sum association Method (WSM)

Madsen and Browning (18) proposed a statistic for testing a pre-specified collapsed set of variants that weights each variant by its frequency, thus allowing one to include variants of any frequency into the collapsed set. This approach proceeds as follows.

Let α be a pre-defined frequency threshold. For the i th variant ($i = 1, 2, \dots, M$), if $p_i < \alpha$, the i th variant is considered as a mutant and its weight is defined as

$$w_i = \sqrt{n_i p_i (1 - p_i)},$$

where frequency p_i of the rare allele A_i is estimated as

$$p_i = \frac{m_i^u + 1}{2(n_i^u + 1)},$$

where m_i^u is the number of mutant alleles observed for variant i in the unaffected individuals, n_i^u is the number of unaffected individuals genotyped for variant i , and n_i is the total number of individuals genotyped for variant i (affected and unaffected). Note that m_i^u takes on values 0, 1, or 2. If $p_i < \alpha$, the weight of i th variant is 0.

The next step is to collapse all the variants which weight is not zero in this region by defining the genetic score of individual j as

$$\gamma_j = \sum_{i=1}^M \frac{I_{ij}}{w_i},$$

where I_{ij} is the number of mutants in variant i for individual j . Thus, for person j , γ_j represents a single score that is obtained by combining information from all the M variants in the region of interest. An association test is performed by testing this score, rather than testing the individual variants. Madsen and Browning (17) suggest using a non-parametric Wilcoxon test for the association test, and calculating the p-value using a permutation approach (see Note 2).

1.6 Pooled Association Tests for Rare Variants

The WSM method described above calculates the weights of the variants for which the frequencies of the rare alleles are less than a predefined threshold α , such as $\alpha = 0.02$ or $\alpha = 0.01$. However, it is difficult to pick an optimal threshold in practice. Price et al. (19) proposed a variable threshold approach for testing rare coding variants.

Variable-Threshold Approach—The idea behind this approach is that there exists some (unknown) threshold T for which variants with a minor allele frequency (MAF) below T are substantially more likely to be functional than are variants with an MAF above T . To obtain this data-driven threshold T , a z-score $z(T)$ for each allele-frequency threshold T is computed, and the maximum z-score across different values of T is defined as $zMax$. A permutation procedure is used to assess the statistical significance of $zMax$, allowing $zMax$ in the permuted data to be attained at values of T different from those in un-permuted data, to ensure the validity of the permutation test. We refer the reader to Price et al. (19) for details about the calculation of the z-scores, and for testing the statistical significance of the variants using this method.

Incorporation of Computational Predictions of Functional Effects—The weights may also be defined based on the functional relevance of the individual variants. Price *et al.* suggest using the PolyPhen-2 scores (20, 21), which evaluate the possible functional effect of a SNP, by calculating the distributions of PolyPhen-2 probabilistic scores for neutral and damaging amino acid changes. The posterior probabilities that a variant is functional are

calculated using these distributions, and used to define the weights. We refer the reader to Price et al. (19) for details about this method.

1.7 Two-Stage haplotype based Methods

Because most of the rare variants are either not genotyped or not well called in GWAS, it is impossible to test specific rare variants or collapsed rare variants. Although the rare variants are usually not well tagged by common variants, it is reasonable to assume that one or more rare variants may fall on only one haplotype consisting of common SNPs. So, another way of exploiting summary statistics for rare variant analysis involves comparing haplotype frequencies between the case and control groups. Zhu et al. (22) have developed a novel association method that can summarize multiple rare variants this way. This two-stage approach proceeds as follows. In the first stage, a set of susceptibility haplotypes are identified by comparing the shared haplotype frequencies in the cases with those in a general population. These susceptibility haplotypes are then compared between cases and controls in the second stage to identify those having significant association with the disease. Although we describe this approach for case-control studies, this method is also applicable to other types of designs such as affected sibpairs.

Suppose we have a total of n individuals, of whom n^u are unaffected (controls), and the remaining $n - n^u$ are affected (cases). In stage 1, randomly sample N ($< n^u$) unaffected and N ($< n - n^u$) affected individuals. Thus, we use the same number of unaffected and affected subjects. We assume that the disease is rare and that the unaffected individuals are representative of the general population. Assume there are N_1 haplotypes shared among the N cases, and that there are k different haplotypes $h_1, h_2, \dots, h_k$ with observed haplotype frequencies $q_1, q_2, \dots, q_k$ among the N_1 shared haplotypes. With the further assumption that the i th haplotype has the corresponding haplotype frequency p_i^0 in the general population, the risk haplotype set is defined as

$$S = \left\{ h_i | p_i - p_i^0 > \gamma \sqrt{\frac{p_i^0(1 - p_i^0)}{N_2}} \right\},$$

where γ is a predefined number that affects the misclassification rate and power. Here $N_2 = 2N$ denotes the total number of individuals used in stage 1.

In the second stage, consider the $n^u - N$ unaffected individuals and the $n - n^u - N$ affected individuals who were not used in the first stage. Further, consider the haplotypes in the risk haplotype set S identified in the first stage. Compare the frequencies of these haplotypes in the affected versus the unaffected individuals considered in the second stage (see Note 3).

2. Methods

We now explain the steps for conducting the two-stage haplotype based analysis of rare variants using the R programming language (<http://cran.r-project.org>). We focus on the

analysis of case-control samples here, although the two-stage method has also been described for sibpairs by Zhu et al. (22). This method involves the following three key steps.

(1) Infer haplotypes (Stage 1 analysis)

We randomly select a subset of cases and controls and infer their haplotypes. For unrelated individuals, software such as Phase, Fastphase and Beagle (24, 25, 26, 27) can be applied. We implemented Beagle to infer haplotypes because of its accuracy and efficiency compared with other software. A description of the Beagle software is available at <http://faculty.washington.edu/browning/beagle/beagle.html>. [The method proposed by Li et al (23) can be used to infer the haplotypes for sibpair data.]

(2) Collapsing the risk haplotypes (Stage 1 analysis)

Once haplotypes are inferred, the risk haplotype set S can be obtained as described above. When grouping the risk haplotypes, we set $\gamma = 1.28$ to control the misclassification rate.

(3) Association test (Stage 2 analysis)

The remaining cases and controls are used to compare the frequency difference in total risk haplotypes between cases and controls using an exact method, such as Fisher's exact test or a test statistic based on an asymptotic approximation.

Example of R code—Suppose that the estimated haplotypes of case-control dataset are stored in a file named *data.txt*. Assume that the total sample size is $n = 5,000$. The first $n - n^u = 2,000$ records are case haplotypes and the remaining $n^u = 3,000$ records are control haplotypes. There is a total of $M = 50$ genetic variants in this region. The data format of file *data.txt* is:

Case_1

14414141432122124324241214241424223222132131111324

14414141432122124324241214241424223222314133311214

Case_2

1423222143424234414423323434212442222332331311224

14232221434244324124233234222142232422314133311214

Case_3

14414141432244324124233234222142232422314133311214

14232221434242344144233234342124422224132131111324

...

Case_2000

14232221434122324124241214241124222212332331331222

14232221434242344144233234342124422224132131111324

Control_1

1423222143424234414423323434212442222332331311224

1423222143424234414423323434212442222332331311224

Control_2

44232223414244324124233234222142232422332331311224

14232221432122124324241214241424223222132131111324

Control_3

44232223414244324124233234222142232422314133311214

44232223414244324124233234222142232422314133311214

...

Control_3000

1423222143412232412424121424112422222132121111324

1423222143424234414423323434212442222332331311224

In this dataset, there are three rows for each individual. The first three rows correspond to person 1 who is a case. Case_1 (the first row) is the case identifier (id). The second and third rows show the inferred haplotypes of this person i.e.,

14414141432122124324241214241424223222132131111324 and

14414141432122124324241214241424223222314133311214. These haplotypes are inferred from Beagle. The remaining rows of this file can be interpreted in a similar manner.

The numbers 1, 2, 3, and 4 in the haplotypes represent the nucleotides A, C, G, and T. The two-stage haplotype based association test can be done using an R code, as described below.

(i) Read the data file, and convert it into a matrix:

```
# read the data

Dt <- scan(filename, what = "character", sep = "\n")

# convert to matrix to matrix

Dtmatrix <- matrix(Dt, ncol = 3, byrow = TRUE)
```

(ii) Randomly select a subset of cases and controls to define the risk haplotypes (Stage 1 analysis), and retain the remaining data for the association test (Stage 2 analysis). In this example, we select 400 cases and 1000 controls for Stage 1:

```
# randomly select 400 haplotype in cases for defining risk haplotypes in the
first stage
```

```
index = sample(1:2000)
```

```
tHap <-as.vector(Dtmat[index[1:400],][,-1])
```

```
# the rest haplotypes in cases will be used in the second stage to do the
association test.
```

```
dHap <-as.vector(Dtmat[index[401:2000],][,-1])
```

```
# randomly sample controls
```

```
index = sample(2001:5000)
```

```
# randomly obtain 1000 haplotype in controls
```

```
sHap <-as.vector(Dtmat[index[1:1000],][,-1])
```

```
# keep the remaining control haplotypes for second stage
```

```
cHap <-as.vector(Dtmat[index[1001:3000],][,-1])
```

```
# create the table of selected haplotypes
```

```
tT <-table(tHap) # 400 haplotypes in cases
```

```
dT <-table(dHap) # 1600 haplotypes in cases
```

```
sT <-table(sHap) # 1000 haplotypes in controls
```

```
cT <-table(cHap) # 2000 haplotypes in controls
```

(iii) Calculate the haplotype frequencies:

```
# calculate the haplotype frequencies
```

```
tHaplotype <- names(tT)
```

```
tHaplotypeFre <- tT/length(tHap)
```

```
dHaplotype <- names(dT)
```



```

dHaplotypeFre <- dT/length(dHap)

cHaplotype <- names(cT)

cHaplotypeFre <- cT/length(cHap)

sHaplotype <- names(sT)

sHaplotypeFre <- sT/length(sHap)T

```

The output of the above steps looks as follows. The variable tHaplotype just contains the list of the haplotypes and the variable tHaplotypeFre is the frequency of each haplotype in tHaplotype:

```

14232221432122124324241214241424223222132131111324  0.00625
14232221434242324122241214342124222212332331331222  0.00250
14232221434242344144233234342124422222332331311224  0.09875

```

(iv) Group the risk haplotypes using the randomly picked 400 haplotypes in cases and 1000 haplotypes in control:

```

tHapMat<-cbind(tHaplotype,tHaplotypeFre)

sHapMat<-cbind(sHaplotype,sHaplotypeFre)

Hapmat<-merge(tHapMat,sHapMat,by.x="tHaplotype",by.y="sHaplotype", all =
TRUE)

Hapmat<-as.matrix(Hapmat)

Hapmat[is.na(Hapmat)]<-0

a1 <- as.numeric(Hapmat[,2]) # number of haplotypes in cases

a2 <- as.numeric(Hapmat[,3]) # number of haplotypes in controls

```

The output of these steps looks as follows. Hapmat consists of three columns corresponding to the haplotypes, the case frequencies of the haplotypes, and the control frequencies. An example is shown below.

tHaplotype	tHaplotypeFre	sHaplotypeFre
14232221432122124324241214241424223222132131111324	0.00625	0.0075
14232221434242324122241214342124222212332331331222	0.0025	0.001
14232221434242344144233234342124422222332331311224	0.09875	0.1185
14232221434242344144233234342124422222314133311214	0.0025	0.0035
14232221434242344144233234342124422222132131111324	0.00125	0.0015
14232221434242344144233234342124422222132121111324	0.00125	5e-04

14232221434242344144233234342124422222134133311214	0.00125	0.001
14232221434242344144233234342124422224132133111324	0.00125	<NA>
14232221434242344144233234342124422224132131111324	0.045	0.037
14232221434242344144233234342124422224132131111222	0.0075	0.0045
14232221434242344144233234342124422224114133331214	0.00125	<NA>

The first line is the title of each column. In the third column, two haplotypes have NA values, meaning that these haplotypes are not found in the controls. For further use, these NA values are replaced with 0 by using the command

```
Hapmat[is.na(Hapmat)]<-0
```

(v) Define the risk haplotypes:

```
ERiskHap <- Hapmat[,1][ (a1-a2) > 1.28*sqrt(a1*(1-a1)/length(tHap) + a2*(1-a2)/length(sHap)) ]
```

If there is at least one risk haplotype (i.e., if

```
length(ERiskHap) > 0
```

), then (**and only then**) run the following commands

(vi) Association testing (Stage 2 analysis). First collect the defined risk haplotypes in cases and in controls for association testing:

```
cHapMat<-cbind(cHaplotype,cHaplotypeFre)
```

```
dHapMat<-cbind(dHaplotype,dHaplotypeFre)
```

```
Hapmat2<-merge(dHapMat,cHapMat,by.x="dHaplotype",by.y="cHaplotype", all = TRUE)
```

```
Hapmat2<-as.matrix(Hapmat2)
```

```
Hapmat2[is.na(Hapmat2)]<-0
```

(vii) Association testing (Stage 2 analysis). Group the risk haplotypes for association testing:

```
pos<-which(Hapmat2[,1]%in%ERiskHap)
```

```
Rmat<-Hapmat2[pos,]
```

(viii) Association testing (Stage 2 analysis). Do the association test using Fisher's exact procedure:

```

# count the number of risk haplotypes in cases

Fd<-sum(as.numeric(Rmat[2,]))

# count the number of risk haplotypes in controls

Fc<-sum(as.numeric(Rmat[3,]))

nd<-length(dHap) # count the total number of haplotypes in cases

nc<-length(cHap) # count the total number of haplotypes in controls

Data = t(matrix(c(Fd * nd, (1 - Fd)* nd, Fc * nc, (1-Fc)* nc), nrow = 2))

Data = abs(Data);

# call the Fisher exact test

Result = fisher.test(Data, alternative = "less")

# calculate p-value

pval_fisher = as.numeric(Result[1])

```

(ix) Association test (Stage 2 analysis). The asymptotic distribution may also be used:

```

stat <- (Fd-Fc)/sqrt(Fd*(1-Fd)/nd+Fc*(1-Fc)/nc) # statistics

pval_asy<-pnorm(stat, mean=0, sd=1, lower.tail = TRUE) # p-value

```

(x) Return the result:

```

return(list(pval_fisher, pval_asy));

```

In our example, the p-values of Fisher's exact test and the asymptotic test were 0.9711 and 0.9994, respectively. Hence, the above code returned the p-value pair (0.9711, 0.9994). We prefer to use the results of the exact test to interpret the result and use the second one as the reference, since some haplotype are rare and, hence, using an exact procedure may be more appealing than the asymptotic approach.

3. Notes

1. The collapsing strategy is based on an important assumption, namely that each individual is unlikely to have more than one rare variant. Given the low frequency of the rare variants, it might be a reasonable assumption. However, the assumption can be violated when the rare variants interact with each other or a large genomic region is considered.

2. A challenge of this method is that it applies a permutation test, which is very time consuming.
3. The haplotype inference can be a challenge if a testing region is large. It may take much computational time for several hundreds of SNPs and a large number of individuals. Also, if many SNPs are considered, it is possible that each individual has his/her own unique haplotypes and this may result in failing to group risk haplotypes. When only unrelated individuals are available, the best approach is to use all the sample in both steps, but then the permutation test should be performed when calculating the p-value (28).

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409(6822): 860–921. [PubMed: 11237011]
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. The sequence of the human genome. *Science*. 2001; 291(5507):1304–51. [PubMed: 11181995]
3. The International HapMap Project. *Nature*. 2003; 426(6968):789–96. [PubMed: 14685227]
4. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007; 449(7164):851–61. [PubMed: 17943122]
5. Chakravarti A. Population genetics--making sense out of sequence. *Nat Genet*. 1999; 21(1 Suppl): 56–60. [PubMed: 9915503]
6. Lander ES. The new genomics: global views of biology. *Science*. 1996; 274(5287):536–9. [PubMed: 8928008]
7. Consortium WTCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447(7145):661–78. [PubMed: 17554300]
8. Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, Sulem P, Thorlacius S, Gylfason A, Steinberg S, et al. Many sequence variants affecting diversity of adult human height. *Nat Genet*. 2008; 40(5):609–15. [PubMed: 18391951]
9. Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C, et al. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet*. 2008; 40(5):584–91. [PubMed: 18391950]
10. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JR, Stevens S, Hall AS, et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet*. 2008; 40(5):575–83. [PubMed: 18391952]
11. Easton DF, et al. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the *BRCA1* and *BRCA2* breast cancer predisposition genes. *Am J Hum Genet*. 2007; 81:873–883. [PubMed: 17924331]
12. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*. 2008; 40(6):695–701. [PubMed: 18509313]
13. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs rare allele hypotheses for complex diseases. *Curr Opin Genet Dev*. 2009; 19:212–19. [PubMed: 19481926]
14. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet*. 2008; 82:100–112. [PubMed: 18179889]
15. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008; 83:311–321. [PubMed: 18691683]

16. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science*. 2008; 322:881–888. [PubMed: 18988837]
17. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res*. 2007; 615:28–56. [PubMed: 17101154]
18. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 2009; 5:e1000384. [PubMed: 19214210]
19. Price AL, et al. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet*. 2010; 86:832–838. [PubMed: 20471002]
20. Ramensky V, Bork P, Sunyaev S. Human Nonsynonymous SNPs: server and survey. *Nucleic Acids Res*. 2002; 30:3894–3900. [PubMed: 12202775]
21. 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; 7:248–249. [PubMed: 20354512]
22. Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol*. 2010; 34:171–187. [PubMed: 19847924]
23. Li X, Chen Y, Li J. Detecting genome-wide haplotype polymorphism by combined use of mendelian constraints and local population structure. *Pac Symp Biocomput*. 2010:348–58. [PubMed: 19908387]
24. Stephens M, Donnelly P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*. 2003; 73:1162–1169. [PubMed: 14574645]
25. Stephens M, Smith N, Donnelly P. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*. 2001; 68:978–989. [PubMed: 11254454]
26. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*. 2006
27. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *Am J Hum Genet*. 2007; 81:1084–1097. [PubMed: 17924348]
28. Feng T, Zhu X. Genome-wide searching of rare genetic variants in WTCCC data. *Human Genetics*. 2010; 128:269–280. [PubMed: 20549515]