NIH Public Access
Author Manuscript
Hum Genet. Author manuscript; available in PMC 2013 September 01.

Published in final edited form as:
Hum Genet. 2012 September ; 131(9): 1395–1401. doi:10.1007/s00439-012-1178-y.

U-statistics in Genetic Association Studies

Hongzhe Li
Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

Abstract

Many common human diseases are complex and are expected to be highly heterogeneous, with multiple causative loci and multiple rare and common variants at some of the causative loci contributing to the risk of these diseases. Data from the genome-wide association studies (GWAS) and metadata such as known gene functions and pathways provide the possibility of identifying genetic variants, genes and pathways that are associated with complex phenotypes. Single-marker based tests have been very successful in identifying thousands of genetic variants for hundreds of complex phenotypes. However, these variants only explain very small percentages of the heritabilities. To account for the locus- and allelic-heterogeneity, gene-based and pathway-based tests can be very useful in the next stage of the analysis of GWAS data. U-statistics, which summarize the genomic similarity between pair of individuals and link the genomic similarity to phenotype similarity, have proved to be very useful for testing the associations between a set of single nucleotide polymorphisms (SNPs) and the phenotypes. Compared to single marker analysis, the advantages afforded by the U-statistics-based methods is large when the number of markers involved is large. We review several formulations of U-statistics in genetic association studies and point out the links of these statistics with other similarity-based tests of genetic association. Finally, potential application of U-statistics in analysis of the next generation sequencing data and rare variants association studies are discussed.

1 Introduction

Analysis of common genetic variants through genome-wide association studies (GWAS) has enjoyed much success over the last few years. More than a thousand genetic loci associated with more than 200 complex traits have been identified (Hindorff et al., 2011). However, these variants typically explain only a small fraction of the inheritable variability for common diseases (Maher, 2008; Manolio et al., 2009). It should be pointed out that almost all these genetic variants were identified by simple single-marker test of association. However, the genetic basis of many common human diseases is expected to be highly heterogeneous, with multiple causative loci and multiple alleles, both rare and common, at some of the causative loci contributing to the risk of these diseases. Analyzing the association of disease with one genetic marker at a time can have weak power due to relatively small genetic effects and the need to correct for multiple testing.

Since multiple-allele within a single gene can be associated with disease phenotype, gene-based tests can potentially provide an important alternative to the simple single SNP test. This has been demonstrated in Huang et al. (2011) and Ngyuen et al. (2010). With the availability of databases of gene functions such as gene ontology (GO) (The Gene Ontology Consortium, 2000) and genetic pathways and networks, it is now possible to test for genetic association between a set of genes with similar functions or a set of genes in the same biological pathways and the phenotypes. This can potentially provide a powerful approach for dealing with genetic heterogeneity (Wang et al., 2007). Both gene-based and pathway-based analyses of GWAS data require tests that can simultaneously take into account multiple SNP markers in a set.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Although testing the simultaneous effects of multiple markers by multivariate statistics might improve power, they will not be very powerful when there are many markers because of the many degrees of freedom. U-statistics (Hoeffding, 1948), which summarize the genomic similarity between pair of individuals and link the genomic similarity to phenotype similarity, have proved to be very useful for testing the associations between a set of SNPs and the phenotype. In general, the U-statistic is defined as the average (across all combinatorial selections of the given size from the full set of observations) of the basic estimator applied to the sub-samples (Lee, 1990). Compared to single marker analysis, the advantages afforded by the U-statistics based methods can potentially be large when the number of markers involved is large. In this paper, we review several formulations of the U-statistics in genetic association studies and point out the links of these statistics with other similarity-based tests of genetic associations, including the distance-based regression (Wessel and Schork, 2006) and kernel machine regression methods (Kwee et al., 2008; Wu et al. 2010).

The advent of next generation sequencing technologies allows one to discover nearly all rare variants in the genome. Testing the aggregated effect of rare variants in a gene on disease susceptibility has become a powerful tool of rare variants association analysis. The idea behind aggregated tests is that if a certain gene is involved in a disease, many rare variants within the gene may disrupt the function of the gene and are therefore associated with the disease. The ideas of the U-statistics can be extended to analysis of rare variants association analysis. We briefly comment on this at the end of this review.

2 Gene- and Pathway-based Tests of Genetic Association - Forming SNP-sets

The motivation behind forming SNP-sets is two-fold. Firstly, it allows us to capture the joint effects of multiple SNPs and harness the linkage disequilibrium (LD) between the SNPs in the SNP-set to increase test power. Secondly, it allows us to incorporate biological information on how SNPs may collectively affect the phenotype of interest, so the results have better biological interpretation. There are various ways to form SNP-sets (see Wu et al. (2010) for an overview). For example, one could form SNP-sets by including all the SNPs that are located near a gene. This could be done by taking all SNPs from the transcription start to end, and possibly including all the SNPs that are upstream and downstream of a gene. A gene-based approach is useful in helping to identify genes that are associated with the disease. Alternatively, for large genes with many SNPs, one can also define the SNP-set based on the LD-block structure of the markers, including the SNPs within a LD-block as the SNP set.

Besides gene-based approach for testing genetic association, pathway- and gene set approaches provide another important alternative. Such approaches aim to test whether the genetic variants in a set of genes with certain biological functions as a whole are associated with disease risk. Such SNP-sets are usually larger than the gene-based SNP-sets.

3 General U-Statistics Formulation for Genetic Association Tests for Case-control Data

Let Y be a random variable that represents the phenotype of interest and **g** be a vector of measured genotypes at K markers. Our goal is to test for association between **g** and Y. We assume that we have a sample of n individuals with phenotype Y_i for the ith individual, where Y_i takes value 0 or 1 in a case-control study and a continuous value in a quantitative

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

trait study. Let $\mathbf{g}_i$ denote a vector of measured genotypes at K markers for subject i, with element $g_{i,k}$ being the kth genotype.

For any two individuals i and j, we can define a score of all genotypes over these K markers using a symmetric kernel $h_{ij} = h(\mathbf{g}_i, \mathbf{g}_j)$. This score or the kernel is usually defined to reflect the similarity in genotypes between these two individuals. A general U-statistic that measures the average score across all pairs of subjects is

$$U_n = \frac{2}{n(n-1)} \sum_{i<j} h(\mathbf{g}_i, \mathbf{g}_j).$$

Let $h_1(\mathbf{g}_i) = E(h(\mathbf{g}_i, \mathbf{G}_j)) = \Sigma_{\mathbf{G}_j} h(\mathbf{g}_i, \mathbf{G}_j) P(\mathbf{G}_j)$, where lower case $\mathbf{g}$ is fixed and uppercase $\mathbf{G}$ is random. Then based on the standard results on the U-statistics, the variance of U_n can be expressed as

$$Var(U_n) = \frac{2}{n(n-1)} [2(n-2)Var(h_1(\mathbf{G}_i)) + Var(h(\mathbf{G}_i, \mathbf{G}_j))]. \quad (1)$$

(see Schaid et al., 2005). Since $Var(h_1(\mathbf{G}_i))$ and $Var(h(\mathbf{G}_i, \mathbf{G}_j))$ only involve summation over the joint genotype $\mathbf{G}_i$ or $(\mathbf{G}_i, \mathbf{G}_j)$, the variance of $Var(U_n)$ can be easily calculated.

To compare the vector of within-group scores for cases with that for controls, one can use the contrast score

$$\delta_{n,m} = \frac{U_{nd} - U_{mc}}{\sqrt{Var(U_{nd}) + Var(U_{mc})}}, \quad (2)$$

where the subscripts d and c denote the diseased cases and controls, respectively. Under the null hypothesis of no differences between cases and controls, standard results for U-statistics imply that $\delta_{n,m}$ has an asymptotic standard normal distribution, i.e., $\delta_{n,m} \rightarrow_d N(0, 1)$.

3.1 The additive kernel function $h(\mathbf{g}_i, \mathbf{g}_j)$

The key of using the case-control contrast U-statistics (2) for testing association between a set of markers and the disease status is to define an appropriate kernel function $h(\mathbf{g}_i, \mathbf{g}_j)$. Schaid et al. (2005) considers a simple kernel function that assumes additivity across all K markers so that $h(\mathbf{g}_i, \mathbf{g}_j) = \sum_{k=1}^{K} w_k h(g_{i,k}, g_{j,k})$, where w_k is the weight associated with the kth marker. Using this additive kernel, the U-statistic U_n can be simplified to

$$U_n = \sum_k w_k \left[\frac{2}{n(n-1)} \sum_{i<j} h(g_{i,k}, g_{j,k}) \right] = \sum_k w_k U_k = \mathbf{w}^T \mathbf{U},$$

where $\mathbf{w} = (w_1, \cdots, w_K)^T$ is the weight vector, $\mathbf{U} = (U_1, \cdots, U_k)^T$ is the vector of marker-specific U-statistic, and U_k is the U-statistic defined for the kth marker. Using this additive kernel, the test statistic contrasting the cases and controls can be written as

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\delta_{n,m}=\frac{\mathbf{w}^T(\mathbf{U}_d-\mathbf{U}_c)}{\sqrt{\mathbf{w}^T\mathbf{V}_0\mathbf{w}}}$$

where $\mathbf{U}_d$ and $\mathbf{U}_c$ are the vector of the U-statistics for the cases and controls respectively, $V_0 = Var(\mathbf{U}_d - \mathbf{U}_c)$ is the variance-covariance matrix of $\mathbf{U}_d - \mathbf{U}_c$. Schaid et al (2005) chose the weight vector $\mathbf{w}$ based on the principal of the best linear unbiased estimator (BLUE), where w_k is proportional to the kth row total of $\mathbf{V}_0^{-1}$.

Schaid et al (2005) provided several possible choices of the kernel function $h(g_{i,k}, g_{j,k})$ for individuals i and j at the maker k, including the allele-match or identify-by-state score, the linear dosage score counting the number of a particular allele. Schaid et al. (2005) conducted simulations and showed that the benefit of using the U-statistic $U_{n,m}$ seemed to occur when there were > 3 high-risk markers among the set of 10 markers in their simulations. In contrast, when there were only one or two high-risk markers, the max-single and multi-marker statistics had greater power. However, it should noted that these results were based on the simulations that assume that the high-risk alleles of the relevant markers were always the minor alleles, i.e., all the minor alleles are risk-alleles.

One problem with using the $U_{n,m}$ statistic with the additive kernel is that when the relevant markers have different effects on disease risk, i.e., some are risk-alleles and some are protective alleles, simply adding the scores across all these markers can potentially lead to elimination of a signal. This has been discussed for the allele-matching kernel. However, this elimination of a signal can also happens for other kernels. This was conformed in Wei et al. (2008). One possible solution to this problem is to first determine the sign of the minor allele effects on disease risk and to take a singed sum of the U-statistic scores across all the markers. However, the null distribution of this modified statistic is not known. One has to use permutation to assess the statistical significance.

3.2 Other possible kernel functions

The contrast statistic (2) is quite general and can be applied to any kernel score functions. Beside the simple allele-matching kernel or IBS-kernel, one can also use the weighted IBS kernel to account for different allele frequencies of the minor alleles of the markers, denoted by q_k for the kth SNP. Then the weighted IBS kernel can be defined as

$$h(\mathbf{g}_i, \mathbf{g}_j)=\frac{\sum_{k=1}^{K} w_k IBS(g_{i,k}, g_{j,k})}{\sum w_k},$$

where $w_k = 1/q_k$ or $w_k = 1/\sqrt{q_k}$. The intuition behind the accommodation of allele frequency is that individuals who share rare alleles may have more similar genomes than do individuals who share common alleles. This kernel provides a possible way of analyzing both common and rare variants together. Beside, the IBS sharing kernel, another promising kernel is the kernel that capture the sharing of identify-by-decent (IBD).

Wessel and Schork (2006) provides several other genomic similarity measures that can be used to define the kernel function and to test for genetic associations, including the IBS allele sharing with weighting for functionality of variations and similarity based on weighting by nucleotide conservation across species. By phasing individuals (i.e., assigning them haplotypes that reflect variations they inherited on their maternally and paternally

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

derived chromosomes), one can assess the similarity of two individuals chromosome pairs by defining unweighted and weighted haplotype-pair similarity.

3.3 An alternative U-statistics based test for case-control data

Following the general idea of Sen (2006), Wei et al (2008) provided another test for association between a set of markers and disease risk. Consider a set of K SNPs. At each SNP, there are three genotypes, coded as $G = 00, 10, 11$. We consider a qualitative trait, taking C different possible categorical values. For example, for case-control studies, there are two trait groups with $C = 2$. Let n_c be the number of individuals in the cth phenotype group. Let $\mathbf{g}_{ci} = (g_{ci,1}, \cdots, g_{ci,K})$ be the observation vector over the K SNPs for the ith individual in the cth group, for $i = 1, \cdots, n_c$, where $g_{ci,k}$ is the genotype of the ith individual in the cth group at the kth SNP that takes one of the three possible genotype values in G. The probability law of $\mathbf{g}_{ic}$ is denoted by $\boldsymbol{\pi}_c = \{\boldsymbol{\pi}_c(\mathbf{g}): \mathbf{g} \in G \times \cdots \times G\}$, where $\boldsymbol{\pi}_c(\mathbf{g})$ is the probability of observing genotype $\mathbf{g}$ in phenotype group c. We are interested in testing the null hypothesis of homogeneity of the $\boldsymbol{\pi}_c$, $c = 1, 2, \cdots, C$.

Since the space of the alternative hypotheses is very large, the standard multi-way contingency table analysis to test for global association suffers loss of power. Instead, following Sen (2006), Wei et al. (2008) considered defining a test statistic based on the U-statistics (Hoeffding, 1948). They first define a symmetric kernel between a pair (i, j) of observations $\mathbf{g}_i = (g_{i,1}, \cdots, g_{i,K})$ and $\mathbf{g}_j = (g_{j,1}, \cdots, g_{j,K})$ as

$$h(\mathbf{g}_i, \mathbf{g}_j) = \sum_{k=1}^{K} w_k I(g_{i,k} \neq g_{j,k}),$$

where w_k is a SNP-specific weight. This kernel function can be regarded as a weighted Hamming distance between individuals i and j over the K SNPs. The definition of this kernel does not depend on particular specifications of the high- or low-risk alleles. Note that here the kernel function $h(\mathbf{g}_i, \mathbf{g}_j)$ measures the differences of the overall genotypes between i and j individuals across K markers, which is different from the kernel scores used in Schaid et al. (2005) that measure the genotype similarity.

Let $n = n_1 + n_2 + \cdots + n_C$ be the total number of individuals across all the C phenotype groups and let U_0 be the pooled group U-statistic corresponding to the same kernel h, which can be written as

$$U_0 = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h(\mathbf{g}_i, \mathbf{g}_j) \tag{3}$$

$$= \sum_{c=1}^{C} \frac{n_c}{n} U_{cc} + \sum_{c \neq c'} \frac{n_c n_{c'}}{n(n-1)} [2U_{cc'} - U_{cc} - U_{c'c'}] \tag{4}$$

$$= W + B, \tag{5}$$

where

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$U_{cc}=\frac{2}{n_c(n_c-1)}\sum_{1\le i<j\le n_c} h(\mathbf{g}_{ci},\mathbf{g}_{cj})$$

is the within-group U-statistics and

$$U_{cc'}=\frac{1}{n_c n_{c'}}\sum_{i=1}^{n_c}\sum_{j=1}^{n_{c'}} h(\mathbf{g}_{ci},\mathbf{g}_{c'j})$$

is the between-group U-statistics. Under the null hypothesis that the genotype distributions over the K makers in the C phenotype groups are the same,

$$B=\sum_{c\neq c'}\frac{n_c n_{c'}}{n(n-1)}[2U_{cc'}-U_{cc}-U_{c'c'}]$$

has zero expectation and it is positive under the alternative.

Wei et al. (2008) proposed to use B/W as a test statistic. However, the distribution of this test statistic is unknown and its significance has to be evaluated through permutations of the phenotypes. They showed that when the minor alleles of the relevant markers are all high-risk alleles, the power of this test is very similar to that of $\delta_{n,m}$. However, the relevant minor alleles include both high-risk and protective alleles, the statistic B/W is much more powerful than the test based on $\delta_{n,m}$.

Alternatively, one can define the following test statistic

$$\delta_b=\frac{B}{\sqrt{Var(B)}},$$

where $Var(B)$ can be calculated using the general theory on U-statistics. Under the null hypothesis, $\delta_b \rightarrow_d N(0,1)$. We would expect similar power of this statistic to that of B/W statistic.

4 *U*-statistics for Quantitative Traits

Several U-statistics-based tests have also be developed for testing association between a set of SNP markers and the quantitative trait phenotypes. Let Y_i be the observed trait value for the ith individual for $i = 1, \cdots, n$. Let $\mathbf{g}_i = (g_{i,1}, \cdots, g_{i,K})$ be the observation genotype vector over the K SNPs for the ith individual for $i = 1, \cdots, n$, where $g_{i,k}$ is the genotype of the ith individual at the kth SNP that takes one of the three possible genotype values $G = \{00, 10, 11\}$, where we assume that allele 1 is the minor allele. The hypothesis that we wish to test is H_0: $F(Y|\mathbf{g}) = H(Y)$, where $F(Y|\mathbf{g})$ is the conditional distribution function of Y given $\mathbf{g}$, and $H(Y)$ is the marginal distribution function of Y.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

4.1 *U*-statistics based on single marker genotypes

To define the *U*-statistics, for marker *k*, Wei et al (2008) defined the set $S_{gk} = \{i: g_{i,k} = g, i = 1, \cdots, n\}$ the individuals with genotype *g* at the *k*th marker for $g \in G$ and $k = 1, \cdots, K$ and let $m_{gk} = |S_{gk}|$ be the number of such individuals. Consider a kernel function between two trait values Y_i and Y_j as

$$\varphi(Y_i, Y_j) = Y_j - Y_i.$$

We define the following *U*-statistics for SNP *k*,

$$\begin{aligned} U_{k1} &= \frac{\sqrt{m_{10k}+m_{11k}}}{m_{10k}m_{11k}} \sum_{i\in S_{10k}, j\in S_{11k}} \varphi(Y_i, Y_j) \\ U_{k2} &= \frac{\sqrt{m_{00k}+m_{11k}}}{m_{00k}m_{11k}} \sum_{i\in S_{00k}, j\in S_{11k}} \varphi(Y_i, Y_j) \\ U_{k3} &= \frac{\sqrt{m_{00k}+m_{10k}}}{m_{00k}m_{10k}} \sum_{i\in S_{00k}, j\in S_{10k}} \varphi(Y_i, Y_j) \end{aligned} \quad (6)$$

which compare the quantitative trait values between every two genotype groups at the SNP *k*. In order to combine these three *U*-statistics, if one assumes that the quantitative trait value is a monotone function of the number of the minor allele at the trait-associated SNPs, one further defines

$$U_k = U_{k1} + U_{k2} + U_{k3}$$

as the *U*-statistic for the *k*the marker. Let $\mathbf{U} = (U_1 \cdots, U_k)^T$ be the vector of the *U*-statistics over the *K* markers. The final test statistics over all *K* markers can be defined as

$$\delta_q = \frac{\mathbf{w}^T \mathbf{U}}{\sqrt{\mathbf{w}^T Cov(\mathbf{U})\mathbf{w}}}, \quad (7)$$

where $Cov(\mathbf{U})$ is the variance-covariance matrix of *U*-statistics vector that can be calculated using the standard technique for *U*-statistics. Following Schaid et al. (2005), the weight vector **w** can be chosen using the BLUP.

4.2 *U*-statistics based on joint marker genotypes

Instead of defining the *U*-statistics for each marker separately as in Wei et al. (2005), Li et al. (2010) defined a similar *U*-statistics based on the joint genotypes over the *K* markers. Suppose the *K* SNPs comprise *L* multi-SNP genotypes, denoted by $G_1, G_2, \cdots, G_L$, where a multi-SNP genotype, G_l, is defined as a vector of *K* single-SNP genotypes that an individual carries. Let's denote by $S_l = \{i: \mathbf{g}_i = G_l\}$ the group of subjects carrying multi-SNP genotype G_l, $l = 1, 2, \cdots, L$ and $m_l = |S_l|$ the number of subjects in group S_l. Li et al. (2010) defined the following global *U*-statistic that measures the overall trait differences among a total of *L* multi-SNP genotype groups,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$U=\frac{\sum_{1\le l<l'\le L} w_{ll'}U_{ll'}}{\sum_{1\le l<l'\le L} w_{ll'}},$$

where

$$U_{ll'}=\sum_{i\in S_l, j\in S_{l'}} \varphi(Y_i, Y_j)$$

is the U-statistic defined for the genotype groups l and l', and $w_{ll'}=\sqrt{m_l+m_{l'}}/(m_l+m_{l'})$ is used to account for the number of subjects in each genotype group. In order to gain power, Li et al (2010) assumes that the expected quantitative trait value of the L multi-SNP genotypes decreases with l (i.e., $E(Y_{S1}) \ge E(Y_{S2} \ge \cdots \ge E(Y_{SL})$. They suggest to sort the multi-SNP genotypes according to their average trait values. Like the case-control contrast U-statistic $\delta_{n,m}$ with the allele matching kernel, when this order of the genotype-effects is misspecified, there is a potential of elimination of signals and dramatically reduced power. In addition, when the data are used to rank these genotypes, the significance of the statistic U has to be evaluated through permutations.

One advantage of such a joint-genotype based U-statistic test is that it can potentially detect epistasis interactions among the set of K SNP markers. One weakness of using the joint genotype is that when K is large, there are a total of 3^K possible genotype combinations, which can be quite large and therefore some sets S_l can be very small. Li et al (2010) proposed to use forward selection and cross-validation for choosing a smaller set of SNPs using the U-statistics as the scores. This can potentially reduce the number of total genotypes in defining the overall global U-statistic.

5 Connection with Genomic Similarity-based Approaches

U-statistics based tests of genetic association are closely related to other similarity-based methods for gene-trait associations. Wessel and Schork (2006) discussed the generalized genomic distance-based regression methods for multilocus association analysis using the distance-based regression methods and psuedo-F statistics of McArdle and Anderson (2001). This approach uses phenotype permutations to assess the statistical significance, which can be time-consuming in genome-wide association analysis.

Tzeng et al (2009) and Tzeng et al (2001) developed gene-trait similarity regression for multi-marker-based association analysis and gene-environment interactions. The key of the gene-trait similarity regression is to regression the pair-wise trait similarity Z_{ij} measure on genomic similarity measure $h(\mathbf{g}_i, \mathbf{g}_j)$:

$$Z_{ij}=b\times h(\mathbf{g}_i, \mathbf{g}_j)+\varepsilon_{ij}, i<j \quad (8)$$

where ε_{ij}'s are some mean-zero error terms, and the trait similarity is defined as

$$Z_{ij}=\{w_i(Y_i-\mu_i^0)\}\{w_j(Y_j-\mu_j^0)\}, \quad (9)$$

which is the weighted cross product of the trait residuals with some weight w_i. The residual is defined with respect to the covariate-adjusted mean for each subject, $\mu_i^0=E(Y_i|X_i)$ under

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the null hypothesis of no association, where X_i is a covariate vector. The weight w_i may be used to account for the fact that Y_i is not necessarily homogeneous. $h(\mathbf{g}_i, \mathbf{g}_j)$ can be defined by any genomic similarity kernel defined above.

The null hypothesis of no association between the K SNP markers and the phenotype is H_0: $b = 0$. Assume that $v_i^0 = var(Y_i|X_i) = m_i^{-1}\varphi(\mu_i^0)$ under the condition of association, where m_i is a known prior weight, such as the binomial denominator, φ is the dispersion parameter, and $v(\mu_i^0)$ is the variance function. This variance function is well-defined for the generalized linear model. Tzeng et al (2009) derive a score statistic for b in model (8),

$$U_b = \sum_{i<j} h(\mathbf{g}_i, \mathbf{g}_j) w_i^{-1} w_j^{-1} (v_i^0)^{-1} (v_j^0)^{-1} (Y_i - \mu_i^0)(Y_j - \mu_j^0). \quad (10)$$

This score statistics is also a U-statistic. Tzeng et al. (2009) further showed the close connection between the score statistics U_b and the score statistic for testing the zero variance ($\tau = 0$) of the random effects in a generalized linear mixed effect model when the covariance matrix of the random effects is specified by τR where R is the genomic similarity matrix.

It is also worth pointing out that if $h(\mathbf{g}_i, \mathbf{g}_j) = \mathbf{g}_i' \mathbf{g}_j$, the score statistic (10) is equivalent to a U-statistic discussed in Zhong and Chen (2011) for testing high-dimensional linear regression coefficients, which uses $\mathbf{g}_i' \mathbf{g}_j (Y_i - \mu_i^0)(Y_j - \mu_j^0)$ as the kernel in the U-statistics. In fact, Zhong and Chen (2011) considered tests for high-dimensional linear regression coefficients for the "large p, small n" situations where the conventional F-test is no longer applicable. They defined a U-statistic

$$T_{n,k} = \frac{1}{P_n^4} \sum^{*} \varphi(i_1, i_2, i_3, i_4),$$

where the summation is over the set $\{i_1 \neq i_2 \neq i_3 \neq i_4$, for $i_1, i_2, i_3, i_4 \in \{1, \cdots, n\}\}$ and $P_n^m = n!/(n-m)!$, and

$$\varphi(i_1, i_2, i_3, i_4) = \frac{1}{4} (\mathbf{g}_{i_1} - \mathbf{g}_{i_2})' (\mathbf{g}_{i_3} - \mathbf{g}_{i_4})(Y_{i_1} - Y_{i_2})(Y_{i_3} - Y_{i_4}).$$

A test of no association between K SNPs and the quantitative phenotype Y can be constructed based on this U-statistic and its asymptotic distribution given in Zhong and Chen (2011).

Another approach for multi-marker association test is the kernel machine regression (KMR) (Kwee et al., 2008; Wu et al. 2010). Since the kernel machine can also be formulated as a generalized linear mixed-effects models, Pan (2010) showed that, when there is no other covariates, if a common positive semi-definite matrix is used as the (centered) similarity matrix in genomic distance regression and as the kernel matrix in KMR, then there is a striking correspondence between the two methods: their test statistics are equal up to some ignorable constants. However, the gene-trait regression and also the kernel regression provide a natural way of incorporating other environmental covariates and possible gene-environment interactions in testing genetic associations.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

One advantage of these genomic similarity-based approaches is that they provide a unified framework for testing genetic association for both binary and quantitative traits. It is however not clear how to unify the statistics reviewed in Sections 3 and 4 for binary and quantitative traits.

6 Discussion

We have reviewed some *U*-statistics that were developed recently for testing the association between a set of SNP markers and the phenotypes, including both the categorial and the quantitative phenotypes. These methods can be applied to gene-based and pathway-based analysis of multi-locus genetic associations and provide useful alternatives to single-SNP based tests for GWAS data. The *U*-statistics based tests are also closely related to distance-based regression and kernel machine regression methods; all these methods focus on relating pair-wise genomic similarity to the phenotype similarity. A comprehensive simulation comparison of statistical powers of these related methods is needed to entangle the subtle differences of the methods.

6.1 Limitations

The expected power gain from testing the association between a set of SNPs or genes and the phenotypes assumes that the SNP set is enriched by phenotype-associated variants. Indeed, if only one genetic variant in a large SNP-set is associated with the phenotype, one should not expect any gain in power when the SNP-set is tested using the U-statistics reviewed above. Including irrelevant SNPs in the SNP-set can certainly lead to loss of power in such set-based tests of genetic association. One possible approach to solve this problem is to assign an important score to each of the SNPs in the set based on their functional relevance or based on the data and then to incorporate such scores into the U-statistics as weights. An interesting approach to obtain such scores is based on the tuned Recursive Elimination of Features (Relief-F) (Moore and White, 2007) or Evaporate Cooling Relief-F (McKinney et al. 2009), an heuristic machine learning methods for estimating the weight of variants. When data-derived scores are used, the test statistics have to be evaluated based on permutations.

6.2 Future Directions

As exon and whole genome sequencing studies are increasingly being conducted to identify rare variants associated with complex traits, methods for analysis of single nucleotide variants from sequencing data are greatly needed (Bansal et al., 2010). Kernel regression method has been developed to test for association between genetic variants (common and rare) in a region and a continuous or dichotomous trait while easily adjusting for covariates (Wu et al., 2011). As reviewed previously, such kernel regression methods are closely related to U-statistics. The kernel method converts genomic information for a pair of individuals to a kernel score representing either similarity or dissimilarity, with the requirement that it must create a positive semidefinite matrix when applied to all pairs of the individuals (Schaid, 2010) and then relates such a genomic similarity to phenotype similarity. Schaid (2010) also provides a comprehensive review of other possible measurements of genomic similarity. Wahba (2012) gave an overview of using dissimilarity data in statistical modeling building. The ideas in Wahba (2012) can be applied to joint analysis of rare and common variants from the sequencing data.

Acknowledgments

This research was supported by NIH grant CA127334. I thank the editor Dr. David-Alexandre Trégouët for inviting me to contribute this review.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

References

Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics. 2010; 11:773785.

The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nature Genetics. 2000; 25:25–29. [PubMed: 10802651]

Hoeffding W. A class of statistics with asymptotically normal distributions. Annals of Statistics. 1948; 19:293325.

Hindorff, LA.; Junkins, HA.; Hall, PN.; Mehta, JP.; Manolio, TA. A catalog of published genome-wide association studies. 2011. Available at: www.genome.gov/gwastudies

Huang H, Chanda P, Alonso A, Bader JS, Arking DE. Gene-based tests of association. PLoS Genetics. 2011; 7:e1002177.10.1371/journal.pgen.1002177 [PubMed: 21829371]

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. American Journal of Human Genetics. 2008; 82:386–397. [PubMed: 18252219]

Lee, AJ. U-Statistics: Theory and Practice. Marcel Dekker; 1990.

Li M, Ye C, Fu W, Elston RC, Lu Q. Detecting genetic interactions for quantitative traits with U-Statistics. Genetic Epidemiology. 2011; 35:457468.

Maher B. Personal genomes: The case of the missing heritability. Nature. 2008; 456:1821.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747753.

McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology. 2001; 82:290297.

McKinney BA, Crowe JE Jr, Guo J, Tian D. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. PLoS Genetics. 2009; 5:e1000432.10.1371/journal.pgen.1000432 [PubMed: 19300503]

Moore, JH.; White, BC. Lecture Notes in Computer Science: Evolutionary Computation, Machine Learning, and Data Mining in Bioinformatics. Springer; 2007. Tuning ReliefF for genome-wide genetic analysis; p. 166175

Nguyen LB, Diskin SJ, Cappasso M, Wang K, Diamond MA, Glessner J, Kim C, Attiyeh EF, Mosse YP, Cole K, Lolascon A, Devoto M, Hakonarson H, Li H, Maris JM. Phenotype restricted genome-wide association study using a gene-centric approach identifies three low-risk neuroblastoma susceptibility loci. PLoS Genetics. 2011; 7(3):e1002026.10.1371/journal.pgen.1002026 [PubMed: 21436895]

Pan W. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. Genetic Epidemiology. 2011; 35:211–216.

Schaid DJ. Genomic similarity and kernel methods II: methods for genomic information. Human Heredity. 2010; 70:132–140. [PubMed: 20606458]

Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN. Nonpara-metric tests of association of multiple genes with human disease. American Journal of Human Genetics. 2005; 76:780793.

Sen PK. Robust statistical inference for high-dimensional data models with application to genomics. Australian Journal of Statistics. 2006; 35:197214.

Tzeng JY, Zhang D, Chang SM, Thomas DC, Davidian M. Gene-trait similarity regression for multimarker-based association analysis. Biometrics. 2009; 65:822–832. [PubMed: 19210740]

Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Bradford BW, Hsu FC, Thomas DC, Sullivan PF. Detecting gene and gene-environment effects of common and uncommon variants on quantitative traits: A marker-set approach using gene-trait similarity regression. American Journal of Human Genetics. 2011; 89:277–288. [PubMed: 21835306]

Wahba, G. Dissimilarity data in statistical model building and machine learning. In: Ji, Lizhen; Poon, Yat Sun; Yang, Lo; Yao, Shing-Tung, editors. AMS/IP Studies in Advanced Mathematics; Fifth International Congress of Chinese Mathematicians; 2012. p. 785-809.

Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. American Journal of Human Genetics. 2007; 81:1278–83. [PubMed: 17966091]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Wei Z, Li M, Rebbeck T, Li H. U-statistics-based tests for multiple genes in genetic association studies. Annals of Human Genetics. 2008; 72:821–833. [PubMed: 18691161]

Wessel J, Schork NJ. Generalized genomic distancebased regression methodology for multilocus association analysis. American Journal of Human Genetics. 2006; 79:792806.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. American Journal Human Genetics. 2010; 86:929–942.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics. 2011; 89:82–93. [PubMed: 21737059]

Zhong PS, Chen SX. Tests for high-dimensional regression coefficients with factorial designs. Journal of American Statistical Association. 2011; 106:260–274.