Genetic Epidemiology 34:222–231 (2010)

Gene, Region and Pathway Level Analyses in Whole-Genome Studies

Omar De la Cruz,[1] Xiaoquan Wen,[1] Baoguan Ke,[1] Minsun Song,[1] and Dan L. Nicolae[1,2]*

[1]*Department of Statistics, The University of Chicago, Chicago, Illinois*
[2]*Department of Medicine, The University of Chicago, Chicago, Illinois*

In the setting of genome-wide association studies, we propose a method for assigning a measure of significance to predefined sets of markers in the genome. The sets can be genes, conserved regions, or groups of genes such as pathways. Using the proposed methods and algorithms, evidence for association between a particular functional unit and a disease status can be obtained not just by the presence of a strong signal from a SNP within it, but also by the combination of several simultaneous weaker signals that are not strongly correlated. This approach has several advantages. First, moderately strong signals from different SNPs are combined to obtain a much stronger signal for the set, therefore increasing power. Second, in combination with methods that provide information on untyped markers, it leads to results that can be readily combined across studies and platforms that might use different SNPs. Third, the results are easy to interpret, since they refer to functional sets of markers that are likely to behave as a unit in their phenotypic effect. Finally, the availability of gene-level *P*-values for association is the first step in developing methods that integrate information from pathways and networks with genome-wide association data, and these can lead to a better understanding of the complex traits genetic architecture. The power of the approach is investigated in simulated and real datasets. Novel Crohn's disease associations are found using the WTCCC data. *Genet. Epidemiol.* 34:222–231, 2010.

Key words: genome-wide association studies; pathway analyses; association in gene networks; combining *P*-values

Additional Supporting Information may be found in the online version of this article.
Contract grant sponsor: NIH; Contract grant numbers: R01 HL087665; R01 DK077489; U01 HL084715.
Omar De la Cruz and Xiaoquan Wen equally contributed to this work.
Omar De la Cruz's present address is Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305.
Minsun Song's present address is Lewis-Sigler Institute, Princeton University, Princeton, New Jersey 08544.
*Correspondence to: Dan L. Nicolae, Departments of Statistics and Medicine, The University of Chicago, 5734 S. University Ave., Chicago, IL 60637. E-mail: nicolae@galton.uchicago.edu
Published online 10 December 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20452

INTRODUCTION

Recent advances in genotyping technology allow the typing of hundreds of thousands of markers at a time (e.g., Affymetrix's Genechip Mapping Array Sets and Illumina Infinium Genotyping Solution), and this has led to a growing number of *genome-wide association studies* (GWAS) [e.g., WTCCC, 2007] in which the whole genome is searched for regions associated with a disease status, often by performing a statistical test of association for each of the typed markers. The GWAS approach is very promising, and has led to many interesting findings of markers showing relatively strong effects. Unless very large sample sizes are available, these studies are not sufficiently powered to find loci of modest effects. The main reason for the lack of power is the multiple-testing adjustment needed to deal with the large number of tests (other reasons are sample ascertainment bias, the potential contribution of rare variants to common disease, imperfect tagging at the platform level and more, but these are not addressed here).

Most of the published GWAS results are based on single marker analyses, where the markers can be genotyped or untyped; for statistical methods dealing with untyped variation see, for example, [Marchini et al., 2007; Nicolae, 2006a,b; Servin and Stephens, 2007]. In general, a method of analysis that treats the different markers as unrelated throws away all the information contained in the structure of the genome. In this article we investigate approaches in which information on the structure of the genome is incorporated by *grouping* the tests. The criterion we use for grouping is that of *functional relatedness*; that is, we group SNPs that are linked to the same functional unit. These functional units can be chosen at least at two levels: basic functional units (e.g., genes, promoter and enhancer regions, etc.), or higher level groupings of those basic units, like metabolic pathways, by considering together the genes and regulatory regions involved in them. In this study we will show examples at the level of the basic units, since databases for pathways are still very incomplete; however, the methods discussed can be applied also at the pathway level, as soon as enough information becomes available.

The goal of this study is not to argue in favor of pathway or gene-level analyses; there are sufficient papers on the advantages and disadvantages of these methods [e.g., Neale and Sham, 2004; Torkamani et al., 2008; Wang et al., 2007]. Our goal is to provide a sound statistical infrastructure for those who choose to do these analyses. For

simplicity, we decided not to use tests for gene-gene interaction in this manuscript, but the ideas presented here can be used directly to construct measures of association that include interaction signals. For example, in a pathway analysis, one can use all marginal association *P*-values and all interaction tests for pairs of SNPs in genes that are directly connected in the pathway.

Gene sets analysis is a well developed topic for studies on differential gene expression [e.g., Goeman and Buhlmann, 2007], but the challenges in those studies are different than those considered here for several reasons. First, the proportion of expected signals is much higher in expression studies. Second, the dependence structure of the data is different and can be better modeled and accounted for in association studies.

Our method is based on the combination of *P*-values obtained from marker-wise analyses. Power is improved by the use of truncation, and a weighting scheme is introduced to deal with correlations due to linkage disequilibrium (LD), as well as to incorporate preexisting information about the markers.

One way to justify our approach is as follows: in a study, if a strong signal is detected for a SNP that lies in a gene, the discovery would be reported as the *gene* being associated with the disease; the same would happen with a known regulatory element; at a higher level, any metabolic pathways, which are affected by the marker, would also become the focus of further studies. One of our goals is to bring attention to functional units in which many weaker signals, considered together, present strong evidence of association for the unit, without losing the ability to detect strong single-marker signals.

As markers become more densely typed, the dependence among the test statistics corresponding to the different markers becomes stronger. Moreover, given the facts that the basic units of genetic function are individually contained in more or less contiguous chromosomal regions, and that the amount of correlation between markers is strongly related to physical distance along the chromosomes, care must be taken when designing a region-specific test statistic [Dudbridge and Koeleman, 2004].

As stated above, we focus this study on contiguous functional elements (e.g., genes), and we call such an element a *region*. The region-specific test statistic that we use is defined using truncation (i.e., using only *P*-values that are small) and relies on weights to incorporate prior information and to deal with the correlation between markers in the region. Other methods have used related strategies, such as splitting the genome into smaller pieces [e.g., Dudbridge and Koeleman, 2004; Gauderman et al., 2007], used truncation [Dudbridge and Koeleman, 2003; Zaykin et al., 2002], or weights [Roeder et al., 2007]. In the Discussion section we describe how our approach is different from those.

The three important issues we would like to emphasize in this manuscript are: (i) possible definitions of the sets—these are described in the next section; (ii) the test statistic and statistical inference; (iii) the issues and complications that arise when interpreting the results obtained using these procedures.

METHODS

THE CHOICE OF SETS AND REGIONS

In this section we discuss possibilities for the type of sets that can be used as input for our inference tools. The following is not an exhaustive list, but gives a good idea on the type of analyses that can be done using the methods we describe in this study.

Genes defined based on physical location. In this situation, we use the sets of SNPs that are within a predefined distance from a gene. For a list of genes, one can use the "UCSC Genes" track from the UCSC Genome Browser [Kent et al., 2002]. These are gene predictions gathered using data from RefSeq, Genbank and UniProt. It is meant to be a moderately conservative set of predictions, in which most sequences are supported by at least two lines of evidence. In this list there are protein-coding as well as putative noncoding transcripts. Each gene region is defined by the starting point (50 Kb before the transcription start point, as given by the UCSC Genome Browser) and the ending point (50 Kb after the transcription end point). The extra 50 Kb regions (total 100 Kb per gene) are included in order to maximize the chance to capture *cis eQTL* variation. There is a trade-off between all possible *cis* eQTLs (that can be anywhere on the chromosome where the gene is located) and keeping the number of SNPs small to reduce the multiple comparison penalty. The choice we make is based on findings from genome-wide expression and variation datasets [Veyrieras et al., 2008]. Notice that we consider only contiguous segments, and therefore include all introns into the corresponding gene region. Also note that gene regions are not necessarily disjoint, for example, due to genes that overlap or are very close in physical location. The main consequence of this is that there are SNPs that contribute to the statistic of more than one gene.

Genes defined based on functional variation. In this case, the set of SNPs corresponding to a gene includes only those that are known to affect the function or expression of the gene. These include: (i) exonic variation such as non-synonymous substitutions; (ii) variation around transcription start site (TSS); (iii) variation around transcription end site (TES); (iv) clusters of transcription factor binding site; (iv) cis and trans eQTLs. Using only functional variation can increase power by decreasing the magnitude of the multiple comparison problem. The drawback is that our databases of functional variation are still incomplete.

Pathways. These are sets of genes together with knowledge about molecular interactions among them. They can be found in databases such as KEGG [Kanehisa and Goto, 2000]. A simple way to analyze the GWA data using pathways is to ignore the interactions within a pathway and treat the genes as exchangeable. This will reduce the pathway information to the set of SNPs that are located within the genes listed in the pathway.

Networks. Networks of genes are designed to reveal information on the structure and function of biological systems. One can model them using directed or undirected graphs, where the nodes are genes and edges represent interactions between genes. We can use the methodology developed in this study to test for enrichment of signals in subnetworks of genes.

Functional regions. At the moment, knowledge about noncoding functional regions is quite incomplete. Although very promising ongoing efforts, like the ENCODE Project [The ENCODE Project Consortium, 2004] aim to find and annotate most regulating regions, at this point they have exhaustively studied only a small part of the human genome [The ENCODE Project Consortium, 2007], which is not very useful for our purpose. Nevertheless, as the databases become more complete,

TABLE I. Summary statistics for the distribution of lengths, in base pairs, of gene regions and conserved regions in Chromosome 22

Type of region	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Gene region	4,025	10,070	20,080	38,340	42,220	883,500
Conserved regions	55	6,708	18,560	44,500	45,160	1,142,000

These segments were obtained from the top level "Chimp net" track in the UCSC Genome Browser (which uses data from their *panTro2* assembly of the chimp genome, from March 2006).

this will become one of the main sources for choosing noncoding functional regions.

Conserved regions. An alternative approach to the one described above is based on comparative genomics. The main idea behind it is to find good candidates for functional elements that can be used as blocks in the genetic analyses. These are regions that are outside of genes (which are known functional units). In this approach, the human genome is aligned and compared to that of other species in order to find conserved regions (segments of chromosomal DNA which are highly similar across species), as those are good candidates to be functional [O'Brien et al., 1999], since it is likely that they were preserved by natural selection. This approach is readily available, due to the fact that the genome of several species, including mammals more or less closely related to humans, has been decoded. We show in Table I summary statistics for segments that are highly conserved between humans and chimpanzee (*Pan troglodytes*), subtracting any intersections these segments might have had with the already selected gene regions. On chromosome 22, approximately 43% of the base pairs are in gene regions, 28% are in the nongene conserved regions and 29.2% lie in regions with no (known) function.

When analyzing the data from a GWAS, one can use a combination of sets; for example, one can select genes and functional regions as sets as well as sets made of individual SNPs that are not included in these functional units.

THE ASSOCIATION STATISTIC: THE CASE OF INDEPENDENT SNPS

The set-specific *P*-values are obtained through a *combined statistic* (CS), which incorporates the evidence provided by each of the SNPs in the region, both genotyped and untyped; this evidence is quantified using the *P*-values obtained for each SNP in a standard marker-wise testing scheme. The CS is then used as a test statistic. Note that, in this approach, possible interactions between SNPs in the same region or pathway are ignored; the focus is on combining the marginal signals for all the variants under investigation. The CS is designed to maximize power against the most likely alternatives. Its construction is based on two concepts: truncation (to increase power) and weighting (to compensate for dependence among markers, and to incorporate prior information). The following two sections describe how we handle these two concepts.

In the case of independent test statistics (corresponding to markers in linkage equilibrium), the CS is obtained by combining the individual marker *P*-values in the region using multiplication, or equivalently, after taking negative logs, using (weighted) addition. This is the same principle used in Fisher's classical method [Fisher, 1932]; however, in order to improve the power of our statistic, we preselect the *P*-values most likely to carry a true signal by using truncation. Truncation has been used in this setting before [Dudbridge and Koeleman, 2003; Zaykin et al., 2002], either by setting a threshold (only *P*-values below the threshold are admitted) or rank truncation (admitting a fixed number of the smallest *P*-values). As a justification for our final choice, we develop here a general framework that includes both approaches as particular cases.

The null hypothesis we want to test is that of no associated markers in the region or set. Assume that there are k markers in the set; then (disregarding the weights for the moment; we will include them later) we define the CS, denoted by C, as:

$$C = -\sum_{i=1}^{k} \log(p_{(i)}) \mathbb{I}_{\{p_{(i)} < \alpha_i\}}.$$

Here $p_{(1)}, \ldots, p_{(k)}$ are the ordered *P*-values, and $\alpha_1, \ldots, \alpha_k$ are the *truncation thresholds*. When all the α_i's are equal, we are in the case of the fixed threshold mentioned above; when $\alpha_1, \ldots, \alpha_r = 1$ and $\alpha_{r+1}, \ldots, \alpha_k = 0$, we obtain the case of rank truncation, admitting only the r smallest *P*-values. When all α_i's are equal to one, we obtain half of Fisher's meta-analysis statistic.

THE TRUNCATION THRESHOLD

The optimal choice of $\alpha = (\alpha_1, \ldots, \alpha_k)$ depends on the underlying genetic model, but there are some characteristics of α that can be learned from a simple investigation of the case of independent test statistics. The alternative hypothesis can be partitioned as $H_A = H_A^1 \cup H_A^2 \cup \cdots \cup H_A^k$, where H_A^i is the hypothesis that there are exactly i SNPs in the region that carry a true signal. It seems intuitively clear that if the truth lies in H_A^r, and if the r associated markers have similar effect sizes that are moderately strong, then we can improve power most by choosing $\alpha_1, \ldots, \alpha_r = 1$ and $\alpha_{r+1}, \ldots, \alpha_k = 0$. Obviously, we do not know in advance how many true signals there are. Nevertheless, since we are testing the alternative that there is *at least* one true signal, it is clear that we should set $\alpha_1 = 1$, that is, always include the minimum *P*-value in our combined statistic. Besides this argument, another advantage of always including the minimum *P*-value is that the distribution of C can be approximated using continuous distributions, while a pure threshold approach yields a distribution with a positive point mass at 0.

The thresholds described in the above paragraph are not useful in practice for several reasons, including the lack of knowledge on the number of expected associations and the similar treatment of genes with different number of SNPs. We describe below two examples of adaptive thresholds that are suggested by procedures used in multiple comparison and model selection problems. As above, we denote with $p_{(1)}, \ldots, p_{(k)}$ the ordered *P*-values from the association tests in the set.

The first threshold is motivated by the idea of controlling the False Discovery Rate (FDR). The original algorithm [Benjamini and Hochberg, 1995] is based on a step-up

procedure where, in order to control for a prespecified FDR value, denoted by γ, one selects the largest j such that $p_{(j)} \leq \frac{j}{k}\gamma$, and reject all $H_{(i)}$ for $i = 1, \ldots, j$. This suggests using $\alpha_i = \frac{i}{k}\gamma$, for a $\gamma \in (0, 1)$ (e.g., $\gamma = 0.25$); this would lead to a procedure that is almost identical to using only the SNPs significant while controlling the FDR at the γ level.

The second scheme is inspired from a procedure used in model selection, namely the higher-criticism [Donoho and Jin, 2004], a procedure that has been suggested for a similar problem in the analysis of gene expression data [Goeman and Buhlmann, 2007]. The higher-criticism threshold is the *P*-value, p_{HC}, that maximizes

$$\frac{i/k - p_{(i)}}{\sqrt{p_{(i)}(1 - p_{(i)})}}.$$

This scheme can be also used in adjusting for multiple comparisons, and has been recently used in analyzing GWA data [Sabatti et al., 2009]. Its application to a multiple comparison setting is based on using all the *P*-values that are below this threshold, and this suggests using the following adaptive threshold for our truncation: $\alpha_i = p_{\mathrm{HC}}$.

Both adaptive thresholds introduced above depend on the number of markers in the set, and aim to include only markers that have a high chance of being associated.

THE TEST STATISTIC: THE GENERAL CASE

The combined statistic defined above assumed exchangeability of the test statistics at the k markers in the set. In practice, this assumption is violated by the nonuniform linkage disequilibrium among markers, by the difference in prior expectation for a marker to be associated (e.g., a nonsynonymous substitution is a better candidate for association than an intronic SNP), and by the difference in the amount of information available (caused by allele frequency and amount of missing data). This suggests using instead,

$$C = -\sum_{i=1}^{k} w_i \log(p_{(i)}) \mathbb{I}_{\{p_{(i)} < \alpha_i\}},$$

where w_i is the weight given to the association test at the ith marker. These weights can be used to incorporate the dependence among test statistics and the previously existing information about the relevance or accuracy of the *P*-values obtained from different markers. The weights are chosen to make the conditional average contributions of log *P*-values match previously chosen values. This can be used in any situation in which we have a priori information on the quality or reliability of the *P*-values arising from different markers; if no such information is available, the weights can be chosen to make the average contributions equal.

The contribution of prior information to the weights is subjective and will not be discussed further in this study, but the programs we have developed allow for user-defined relative weights for all the markers in the human genome. These priors can contain information on the relative risk of a marker to be functional: for example, we can up-weight the results coming from nonsynonymous substitutions. The contribution of the amount of available information to the weights can be easily quantified using asymptotic relative efficiency [Nicolae, 2006a,b], and the case of untyped markers is discussed in the next section.

The most difficult inference is on determining the LD contribution to weights, and we use the rest of this section to discuss our approach for this problem.

We use weights with the guiding principle that, under the null hypothesis, all markers should have the same (average) contribution to the combined statistic, conditionally on it actually being used (that is, the corresponding *P*-value being the minimum among the k SNPs in the region, or it being smaller than the threshold α). The reason for this is the following: if a few markers are in linkage disequilibrium, the corresponding *P*-values will tend to be positively correlated (when using two-sided tests as it is the norm in association studies). Then, under the null, if one of them is small by chance, the others will tend to be small too, leading to an artificially large combined statistic. By applying lower weights to groups of markers whose *P*-values are positively correlated, conditional on being small, we compensate their inflating effect without affecting the statistic in the case when the *P*-values are not so strong (since then they are simply discarded). This criterion is equivalent to making the covariance between C and each of its components equal. In the extreme case, when several markers are in perfect LD, they are treated as a block (all of them contribute to C, and by the same amount, or none does), and therefore are weighted down so that all together contribute as much as a single SNP. Most cases are not so extreme, but the same principle applies. This method provides a way of dealing with dependence that is relatively inexpensive computationally. Note that this is in contrast with other approaches [Eskin, 2008] that give higher weights to markers that are proxies for many typed or untyped variants.

Explicit computation of the weights might be possible (though cumbersome) if the full joint distribution of the *P*-values under the null hypothesis is known. In the GWA applications we will not have that information, and must rely on the use of permutations; this, however, also presents some difficulties. In theory, instead of computing empirical correlations from simulations or permutation schemes, we can use permutations to compute the weights directly, as follows: out of the resulting *P*-values from a large number of permutations, for each marker one restricts the computation of the expected value of C to those permutations which make the *P*-value for that marker small enough to be admitted in the truncation step.

The problem with this scheme is that in order to obtain enough cases in which the *P*-value is admitted so that the conditional average contribution of the marker can be computed reliably, we need to perform a very large number of permutations. Since this needs to be repeated for each region (the number of which can be in the order of tens of thousands), this approach becomes unfeasible, especially when the thresholds in the truncation step are very small (as we will need them to be).

To circumvent this problem we compute instead the conditional contributions by conditioning with a larger threshold λ, as an approximation of the actual conditional contributions. For that we define

$$C_\lambda = \sum_{i=1}^{k} w_i X_i \mathbb{I}_{\{p_i < \lambda \text{ or } p_i = p_{(1)}\}}.$$

We will use $C = C_\alpha$ as our statistic, but in order to find the weights we will perform permutations with $\lambda > \alpha$. Each

weight w_i is computed so that $E[C_\lambda|p_i<\lambda]$ is the same for all $i=1,\ldots,k$. Or, in the case where we have pre-specified contributions v_i, $i=1,\ldots,k$, we choose the weights so that the contributions $E[C_\lambda|p_i<\lambda]$ are proportional to the v_i's. See the Supplementary Materials for the actual algorithm. As we use it here, λ is a tuning parameter that, for a fixed number of permutations, adjusts the trade-off between bias and variance for the estimation of the weights. Thus, λ can be made smaller to reduce bias, but with increased computational costs.

UNTYPED MARKERS

The CS described above includes tests done at both genotyped and untyped markers. Generally, there are two main groups of methods for testing untyped variation. The first group includes methods that use cluster-based models for population genetic variation, with cluster membership estimated using Hidden Markov Models along entire chromosomes [e.g., Marchini et al., 2007; Servin and Stephens, 2007]. The second group of methods use simple predictors for untyped allele frequencies, with the training done in reference databases such as HapMap [e.g., Nicolae, 2006a,b]. The result of using the imputation methods is a set of *P*-values, for each of the typed markers and for each of those untyped markers for which there is enough information to perform the test.

Our approach for untyped variation (TUNA) is described elsewhere [Nicolae, 2006a,b; Wen and Nicolae, 2008], and relies on estimates of joint haplotype frequencies for both typed and untyped markers in order to perform tests of association for the untyped markers. The frequencies are used to build linear combinations of observed haplotype frequencies that are used as proxies for untyped alleles. The method also provides a measure of quality for the untyped markers, M_D [Nicolae, 2006a,b], which quantifies the amount of information provided by nearby typed markers.

Our weighting scheme is very useful in this situation. We define a set of relative information weights (desired contributions) by giving maximum weight to the typed markers, and a reduced weight for each imputed SNP, based on the amount of available information for that marker. The actual weights are computed so that the conditional average contribution of each marker to the CS is indicative of the power to detect association using that marker. The relative weights we propose are equal to M_D, which is defined as the asymptotic relative efficiency for the observed data versus the complete data (as if the marker was genotyped), i.e., the relative sample size available for the imputed marker. Because the χ^2 association statistic has the same order of magnitude as the sample size, this choice of weights will lead to equal contributions of the noncentrality parameters when the effect sizes are equal.

THE REGION/SET *P*-VALUES AND THEIR INTERPRETATION

The combined, set-specific *P*-value is obtained by comparing the value of the combined statistic *C*, obtained from using the true cases and controls, to the distribution of *C* under the null hypothesis of no association for any of the SNPs in the particular set. There are two ways one can obtain the null distribution: empirically from simulations, or using statistical modeling possibly coupled with large sample approximations. When using *C* for independent markers, it is possible to derive approximations to the null distribution; in a simpler scenario, the generalized extreme-value distribution has been used [Dudbridge and Koeleman, 2004]. This task is more difficult under most practical scenarios involving strong LD between markers. That is why, in our implementation, the null distribution is obtained empirically using permutations in which individuals are assigned to ''case'' and ''control'' groups at random. Note that the same permutations used in estimating the weights are used for obtaining empirical distributions for the test statistics, and these distributions are region-specific because of the difference in the number of markers and LD patterns.

Note that the *P*-values obtained as above are adjusted for the within-set multiplicity problem, i.e., they have a uniform (0,1) distribution under the null hypothesis that there are no associated SNPs in the set. The main issue in their interpretation is that the test statistics and their corresponding *P*-values are not exchangeable under the alternative hypothesis. For an example of this, suppose identical *P*-values are obtained from two sets: one with one genotyped SNP, and second with 100 SNPs out of which only 50 are genotyped. This problem is not unique to this situation, and we encounter it even in cases where interpretation seems straightforward. For example, we rank SNPs based on their *P*-values even when they have different allele frequencies (and some SNPs might be ranked higher than others even if they have lower odds ratios), or when they are based on different information such as sample size (when ranking genotyped and imputed SNPs together). Bayesian solutions are possible when modeling the alternative is straightforward, but this is not the case in this situation. Our recommendation for this case is against ranking, and just for using the *P*-values as measures of evidence for/against the presence of association in the tested sets/regions.

SIMULATED DATA

We performed diverse simulations to evaluate the performance of the methods discussed here; these simulations were divided in two groups. In the first group we compared different thresholding schemes; this allowed us to select the best performing scheme, and at the same time compare it with other existing approaches. In the second group we compare the performance of our method, now including the use of weights, with a method using principal components analysis (PCA) [Gauderman et al., 2007].

SIMULATIONS: THRESHOLDING

We performed simulations to determine which form of adaptive thresholding provides highest power in a consistent way against a wide array of reasonable alternative hypotheses and significance levels. These simulations confirmed that always including the minimum *P*-value in the CS results in an increase in power; after that, only the rest of the threshold needed to be determined. The behavior of the threshold points $\alpha_2,\ldots,\alpha_k$, corresponding to $p_{(2)},\ldots,p_{(k)}$, can be roughly classified in three groups: increasing, decreasing and constant.

The first scheme, described earlier, motivated by the step-up FDR procedure, leads to a linearly increasing behavior. Rank methods that admit a prespecified number of the lowest *P*-values show a decreasing behavior; other decreasing schemes, like exponential decay, behaved similarly in our simulations. The second scheme in the earlier section, based on the higher-criticism procedure, is somewhat different, since the number of admitted *P*-values depends on the whole set of *P*-values. For simplicity, we used instead a variation on this method, in which $p_{(i)}$ is admitted into the CS if its standardized deviation from the expected value under H_0 is above a fixed threshold (say, a quantile for the standard normal distribution); this leads to an increasing sequence of threshold points, but with upward concavity.

The artificial data for this set of simulations was created at the level of the *P*-values. These were generated from the uniform distribution, for the null case, and for the alternative by finding the right-tail probability, under the central χ^2 distribution with 1 degree of freedom, of draws from a noncentral χ^2 distribution with 1 degree of freedom. The combined statistics where obtained using the different schemes, and the null distribution was obtained by using 5,000 replications under the null. With this distribution, the power was computed, at different significance levels, for different noncentrality parameters (mostly in the range from 1 to 10). Other factors that were varied were the number of SNPs in the region having a true alternative distribution (between 1 and 10), and the total number of SNPs in the region (in the order of 100).

Of all the thresholding schemes described above, the most consistent good performance was obtained for constant thresholds (after admitting the smallest *P*-value), when the constant threshold is chosen to be of the magnitude of the 0.005 quantile of the distribution of the second smallest *P*-value, under the null hypothesis (for example, if there are $k = 100$ SNPs in the region, and there is no dependence, then the threshold comes to be approximately 0.00104). However, we have chosen to estimate instead the 0.1 quantile of the distribution of the smallest *P*-value, which can be estimated more accurately with a given number of permutations.

Increasing schemes performed worse, in general. Decreasing schemes, like fixed rank thresholds, outperformed the constant threshold when the chosen rank matches the true number of alternatives, but perform worse overall.

The results of some of the simulations are plotted in Figure 1. The plots show the power of tests based on the different thresholding schemes, depending on the significance level used, the number of true signals and, in the case of constant thresholds (but including the minimum), the actual level of the threshold.

SIMULATIONS: WEIGHTS

Once settled in the thresholding scheme, we compare our weighting scheme with the PCA-based method in Gauderman et al. [2007]. We performed the computations as described [Gauderman et al., 2007], choosing enough components to guarantee that at least 80% of the variation is explained, as was done in their simulations.

In this second group of simulations we generated the artificial data at the level of genotypes, creating data sets that resemble the actual data that would be used with these methods. This was done as follows: the individuals are divided in two equal sets, cases and controls. The markers (which make up a region) are divided in several blocks; for each block, the first marker is chosen as the "leader." The genotype of each leader marker is drawn from a binomial distribution with $n = 2$, independently for each individual, using a pre-specified probability for cases and another for controls (these may vary from block to block). Then the genotypes for the other markers in each block are derived from the genotypes of the leader marker by successively making random changes based on pre-specified transition probabilities that quantify the probability of change for each allele (these also may vary from block to block). By setting the binomial probabilities for cases and controls we affect the strength of association with disease status (for the leader marker; the association weakens for the other markers as correlation with the leader decreases), and by setting the transition probabilities we affect the level of correlation between markers inside a block. Markers in different blocks are independent.

For each artificial data set we obtain a *P*-value using our method, by first obtaining a *P*-value for each marker using logistic regression—this is equivalent to using Armitage's trend test [Sasieni, 1997], and then combining the *P*-values as described in the study, with weights calculated as described in Supplementary Materials. Another *P*-value is obtained using the method of Gauderman et al. [2007]; this allows us to compare the two methods side by side. The result of these simulations is that our method is able to detect weak signals in situations where the method of Gauderman et al. is not; in particular, when the number of markers in the region is in the order of 100 (the advantage of our method vanishes when the region contains a number of markers in the order of 10).

As an example, we describe here a typical simulation run: 15 independent simulations were run according to the description above, with 400 cases and 400 controls, and 100 markers in the region, divided in two blocks of 50. The leader of the first block is associated with the disease status, with a probability parameter of 0.3 in the cases and 0.4 in the controls; the leader of the second block is not associated with the disease, having a probability parameter of 0.3 for both cases and controls. In the first block, the probability of changing the genotype by one allele is 0.2 from marker to marker; in the second block, the probability is 0.01. In other words, in the 100 markers only the leader of the first block, and a few of the nearby markers, have any association with the disease status. For each simulation, *P*-values were computed using the PCA method as well as the method described in this study. The resulting *P*-values are listed in Table II; clearly, our method detects the signal, while the PCA method does not. Larger sample sizes lead to smaller *P*-values, and require more permutations (here we used 15,000).

APPLICATION: CROHN'S DISEASE

As an illustration of our method, in this section we apply it to data obtained from a case-control study of Crohn's disease, namely the Wellcome Trust Case-Control Consortium [WTCCC, 2007]. In this study, DNA samples from 2,000 cases and 3,000 controls were genotyped using Affymetrix 500K SNP array. Both individuals and SNPs that did not pass quality control were removed under the guidelines provided by the main WTCCC paper [WTCCC, 2007]. The final dataset consists of 2,938 control and 1,748

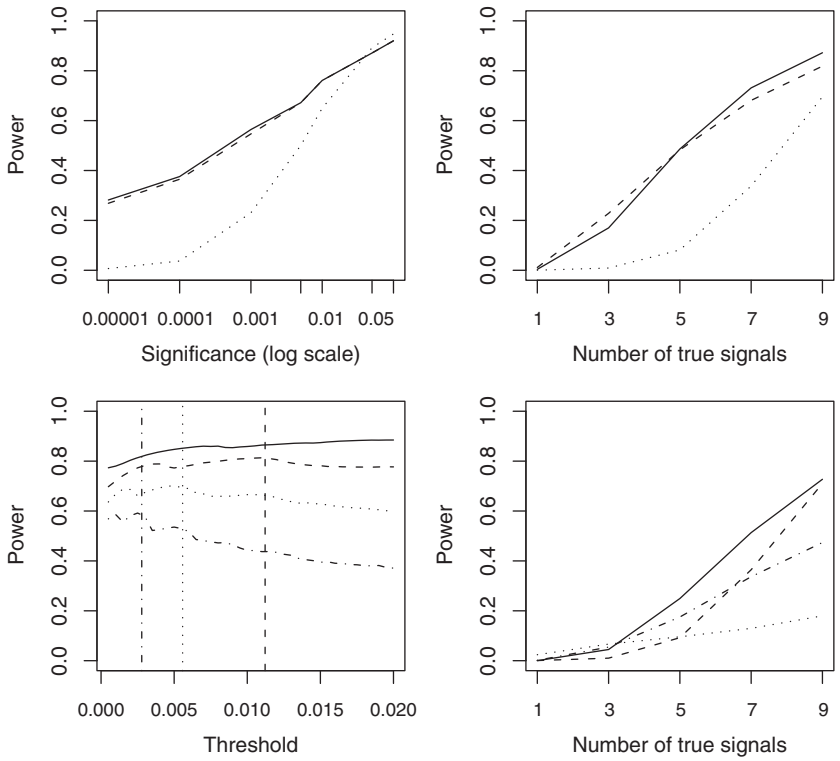

Fig. 1. Plots of power under different situations. Top left: Here we compare three ways to define the ordered thresholds, all of them always accepting the minimum *P*-value. Solid line: constant threshold, with value 0.001. Dashed line: thresholds decreasing exponentially. Dotted line: thresholds increasing linearly from 0.001 to 0.5. The constant threshold performs as well or better than the others. The number of true alternatives is 4, out of a total of 100 markers. Top right: We compare again the three methods from the first plot, now across several numbers of true signals among the markers in the region, using a significance level of 0.0001. Bottom left: Using constant thresholds (after including the minimum), we compare the performance of different threshold sizes. Solid line: The region contains 16 markers. Dashed line: The region contains 32 markers. Dotted line: 64 markers. Dash/dot line: 128 markers. The vertical lines of the same type stand at the 5% quantile for the second smallest *P*-value, assuming the corresponding markers are independent (that is, the 5% quantile of the Beta(2, $n-1$) distribution, where n is the number of markers in the region). This plot suggests that our choice for the threshold level is reasonable. The number of true alternatives is kept constant at 3, and the significance level at 0.0001. Bottom right: Here we compare the performance of our thresholding scheme (always including minimum, then constant threshold at 0.001) against other schemes. Solid line: our method. Dashed line: plain threshold (that is, not forcing the minimum in), at value 0.1. Dotted line: Minimum *P*-value. Dash/dot line: Rank truncation, keeping the five smallest *P*-values. Our method performs better across several numbers of true signals, out of a total of 100 markers.

case samples. A total number of 397,733 autosomal SNPs with MAF >0.01 within the WTCCC Crohn's study are included in the final dataset for analysis. The single SNP analysis duplicates the association results reported in the original WTCCC paper, so we are confident that we are working on the same set of subjects and markers.

We obtained gene annotation from the RefSeq gene table for human genome build 35 (hg17) [Pruitt et al., 2007]. As mentioned in method sections, we define a particular gene set as SNPs in the region from 50 kb upstream of the gene transcription starting site to 50 kb downstream of the gene transcription ending site. By this definition, the filtered SNP set covers 22,079 of all 27,222 possible annotated autosomal genes.

We performed the proposed procedure on all covered genes with significance and weights obtained using 10,000 permutations. The region *P*-value for each gene is computed using the optimal weights described above, and the threshold that is determined by the 10th percentile of the empirical distribution of minimum *P*-value for the given gene SNP-set. This analysis identified 51 genes with combined *P*-value $<1\times10^{-4}$ (i.e., no permutation showed a combined statistics that is as large as the observed one). For this set of genes, we further performed 1 million permutations to get more accurate *P*-values. Note that out of these 51 genes, 32 have the minimum single-SNP *P*-values larger than 1×10^{-5} (this is the threshold used for reporting in the WTCCC paper), so these can correspond to findings not reported in the original analysis. The number of false positives among these is low—we expect, on average, only 2–3 out of them to be false discoveries.

There are 11 genes with *P*-values smaller than 1×10^{-6}. They correspond to regions that are discussed in the WTCCC analysis, and are not further investigated here. As a side note, some of the very significant genes, e.g., CYLD on chromosome 16, are very likely associated only indirectly as their designated regions contain markers within or in LD with variation in disease genes (NOD2 in the case of CYLD). Our focus in this study is different, and, more interestingly, we are able to identify genes showing

TABLE II. Resulting *P*-values from using the PCA method [Gauderman et al., 2007] and from the method described in this study ("region *P*-value"), using 15 simulated data sets with few true signals in a region containing 100 markers, with correlation among markers (see details in the text)

PCA-based *P*-value	Region *P*-value
0.15001	0.00147
0.70052	0.00587
0.05918	0.00100
0.24100	0.00001
0.15378	0.01620
0.68219	0.02273
0.84839	0.00033
0.48214	0.00087
0.72955	0.00060
0.44948	0.13667
0.63074	0.00067
0.29848	0.00440
0.92574	0.01107
0.03492	0.01460
0.30763	0.01307

multiple independent modest association signals. We illustrate some of them in Figure 2. In panel A of Figure 2 we show the results for a highly interesting candidate gene, IRF8 on chromosome 16q. The length of this gene is 23.4 kb. The smallest *P*-value in this region is equal to 0.0004 (it is 0.01 after adjusting for multiple testing within the region). The combined *P*-value (0.0001) suggests multiple independent signals in this gene. A similar story emerges for ICOSLG (panel B, length 14.1 kb) on chromosome 21q (combined $P = 5 \times 10^{-5}$), a gene that has been already confirmed as a Crohn's gene using a meta-analysis [Barrett et al., 1995]. Finally, in panel C, we show the results for a pseudo-gene, ATQL4 ($P = 2 \times 10^{-6}$) of length 449 bp, on chromosome 10q21 in a region that is mentioned in WTCCC [2007]. Note that this gene is genome-wide significant even after a Bonferroni correction (using the number of genes, 22,079), and that none of markers in the 10q21 region that are highly significant in the single-marker analysis (e.g., rs10761659) are present in the set used for the region analysis.

DISCUSSION

We propose a novel method for analyzing data from genome-wide association studies, based on combining markers into *sets*, blocks of the genome, which are selected so that they correspond to functional units such as genes, regulatory regions, or pathways. This way, the grouped markers are more likely to be associated to the same phenotype.

The joining of the information from different markers in the same set is achieved by a meta-analytic approach in which the *P*-values from individual-marker tests are combined into a statistic related to Fisher's classic method, but improved through the use of truncation, which increases the power of the resulting test by incorporating only *P*-values likely to carry information, and of weights, which are used to compensate for the effect of the correlations that often exist among the markers in a contiguous genomic region, and to include prior information on the SNPs in the set.

Our method provides a powerful way to identify functional sets, which are associated to a disease status. The increase in power when compared to single-marker approaches comes from the fact that moderately strong signals can be combined into a stronger signal. Also, in combination with the methods for imputing untyped variation, it generates results that can be compared easily across different studies, even on different platforms, since the corresponding set-specific *P*-values refer to tests of the same hypotheses. Finally, the results from these analyses offer a clearer picture on the functional relevance of the discoveries.

Some elements of our method have been used by other researchers; here we describe how our strategy differs from previous ones, both in philosophy and detail. The strategy of grouping markers has been used often before. However, this has been done based on the linkage disequilibrium between the markers, usually by considering haplotypes, either by using actual haplotype information [Morris and Kaplan, 2002] or genotype information [Epstein and Satten, 2003], but restricted to areas of high LD (haplotype blocks). Other ways of grouping markers based on LD use clustering [Durrant et al., 2004] or principal components [Gauderman et al., 2007]. Unlike these, we group markers based on functional considerations, which we believe is in better accordance with the goal of identifying variation that is associated with phenotype. Even if some studies are designed to use variation within a functional unit, like a particular gene, in our case we use this idea in a genome wide setting. Furthermore, our approach can be easily extended to higher level functional units, like pathways, once enough pathway information becomes available.

The use of truncation in the meta-analytic approach for combining information from different markers has been used before, either by setting a threshold [Zaykin et al., 2002], or by picking a fixed number K of the smallest *P*-values [Dudbridge and Koeleman, 2003] (this is called *rank truncation*; it includes as a special case the method of combining *P*-values which simply chooses the minimum). In this article we investigated both approaches, in a more general setup in which we allow a different threshold for each *P*-value, dependent only on the rank of the *P*-value. Our conclusion is that the most consistent performance, across a wide range of alternative hypotheses, is obtained by always including the minimum *P*-value, and truncating the rest with a stringent threshold. This approach does not weaken the signal from a single strong *P*-value (as can happen with the pure threshold approach), but allows the combination of several somewhat weaker signals from the same region into a strong region specific *P*-value, without need to search through several values of K, as suggested for rank truncation [Dudbridge and Koeleman, 2003]. In particular, our simulations show that the method proposed here outperforms using the minimum *P*-value, which has been proposed as a reasonable summary for tests in a region [Kaplan and Morris, 2001].

Weighting schemes have been proposed as a way to introduce previously existing information about the markers. These weighted procedures multiply the threshold by the weight w, for each test, raising the threshold when $w > 1$ and lowering it when $w < 1$ [Genovese et al., 2006; Roeder et al., 2002; in Roeder et al., 2002, grouping is also considered, but only as a way to manage the weights]. An alternative strategy

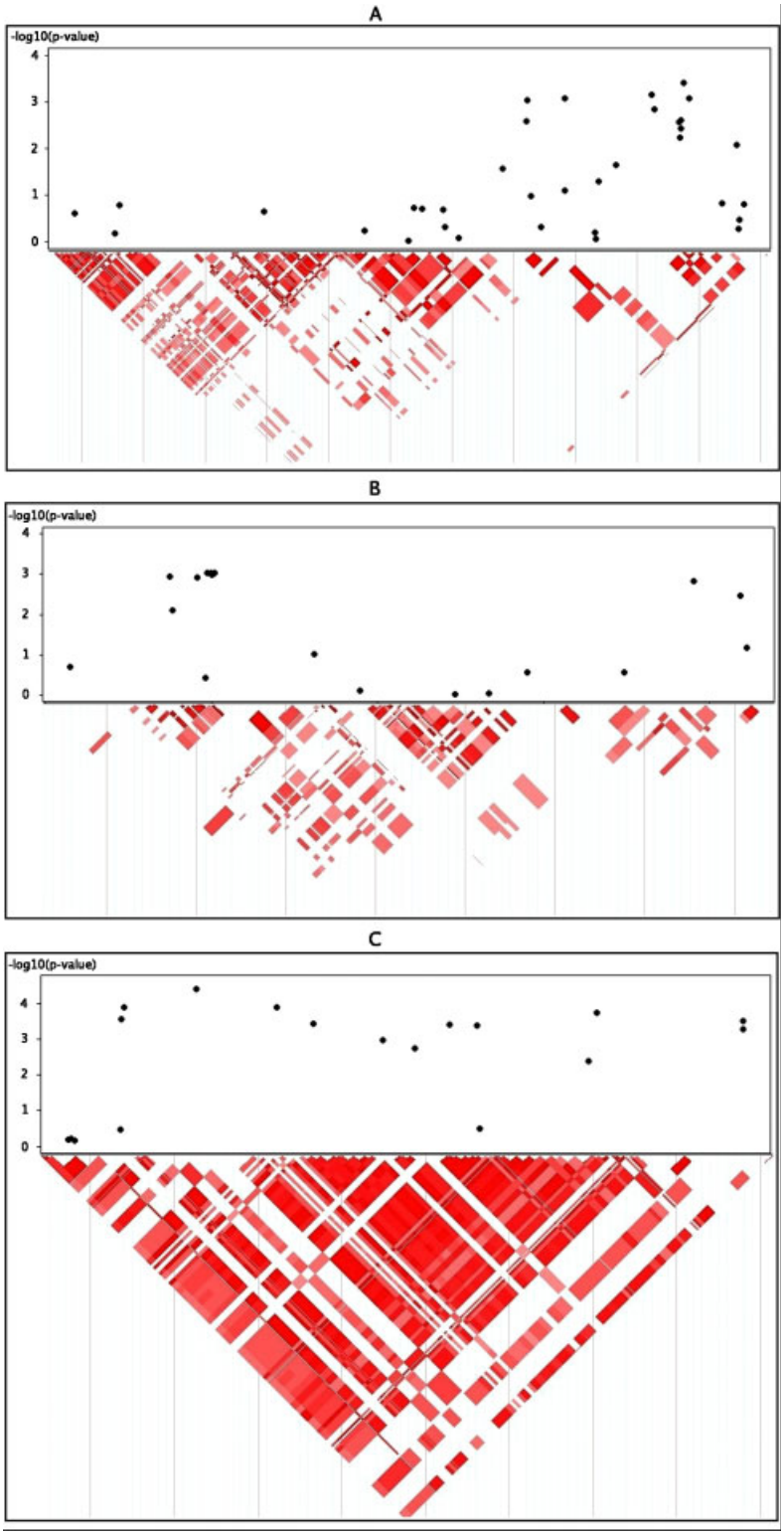

Fig. 2. Association *P*-values (on $-\log_{10}$ scale) and the r^2 measure of linkage disequilibrium are plotted for three genes: IRF8 (A), ICOSLG (B), ATQL4 (C).

is the use of strata that use prior information on the probability of a marker to be associated [Sun et al., 2006]. In our case, the weights are used only inside each functional region, and even when they are used to incorporate previous information, one of their main roles is to help compensate for the effects of dependence. Furthermore, our use of the

weights with the previous information on marker quality obtained from TUNA is novel.

Extensions of this method are in progress and include: (i) incorporating gene-gene interaction signals for the markers within the same region or pathway; (ii) updating the prior information used in the individual weights by incorporating relevant information from SNP annotation studies; (iii) generalizations to incorporating rare variation that is obtained from re-sequencing studies.

ELECTRONIC REFERENCES

We have started to implement the methods developed in this study in the software package SLAT (Set Level Association Testing): http://www.stat.uchicago.edu/~wen/slat/.

ACKNOWLEDGMENTS

We are grateful to the Wellcome Trust Case Control Consortium for sharing their GWA data with us.

REFERENCES

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ; NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. 1995. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40:955–962.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. J R Stat Soc B 1:289–300.

Donoho D, Jin J. 2004. Higher criticism for detecting sparse heterogeneous mixtures. Ann Stat 32:962–994.

Dudbridge F, Koeleman BPC. 2003. Rank truncated product of *P*-values, with application to genomewide association scans. Genet Epidemiol 25:360–366.

Dudbridge F, Koeleman BPC. 2004. Efficient computation of significance levels for multiple associations in large studies of correlated data, including Genomewide Association Studies. Am J Hum Genet 75:424–435.

Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. 2004. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 75:35–43.

Epstein MP, Satten GA. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73:1316–1329.

Eskin E. 2008. Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. Genome Res 18:653–660.

Fisher RA. 1932. Statistical Methods for Research Workers. London: Oliver and Boyd.

Gauderman WJ, Murcray C, Gilliland F, Conti DV. 2007. Testing association between disease and multiple SNPs in a candidate gene. Genet Epidemiol 31:383–395.

Genovese CR, Roeder K, Wasserman L. 2006. False discovery control with *P*-value weighting. Biometrika 93:509–524.

Goeman JJ, Buhlmann P. 2007. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23:980–987.

Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30.

Kaplan N, Morris R. 2001. Prospects for association-based fine mapping of a susceptibility gene for a complex disease. Theor Popul Biol 60:181–191.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The Human Genome Browser at UCSC. Genome Res 12:996–1006.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913.

Morris RW, Kaplan NL. 2002. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221–233.

Neale BM, Sham PC. 2004. The future of association studies: gene-based analysis and replication. Am J Hum Genet 75:353–362.

Nicolae DL. 2006a. Testing untyped alleles (TUNA)—applications to genome-wide association studies. Genet Epidemiol 30:718–727.

Nicolae DL. 2006b. Quantifying the amount of missing information in genetic association studies. Genet Epidemiol 30:703–717.

O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, Copeland NG, Jenkins NA, Womack JE, Marshall Graves JA. 1999. The promise of comparative genomics in mammals. Science 286:458–481.

Pruitt K, Tatusova T, Maglott D. 2007. NCBI reference sequences (Ref-Seq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–D65.

Roeder K, Devlin B, Wasserman L. 2007. Improving power in genome-wide association studies: weights tip the scale. Genet Epidemiol 31:741–747.

Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, Jones CG, Zaitlen NA, Varilo T, Kaakinen M, Sovio U, Ruokonen A, Laitinen J, Jakkula E, Coin L, Hoggart C, Collins A, Turunen H, Gabriel S, Elliot P, McCarthy MI, Daly MJ, Jarvelin MR, Freimer NB, Peltonen L. 2009. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet 41:35–46.

Sasieni PD. 1997. From genotypes to genes: doubling the sample size. Biometrics 53:1253–1261.

Servin B, Stephens M. 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3:e114.

Sun L, Craiu RV, Paterson AD, Bull SB. 2006. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. Genet Epidemiol 30:519–530.

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia of DNA Elements) project. Science 306:636–640.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816.

Torkamani A, Topol EJ, Schork NJ. 2008. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics 92:265–272.

Veyrieras B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4:e1000214.

Wang K, Li M, Bucan M. 2007. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 81:1278–1283.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Wen X, Nicolae DL. 2008. Association studies for untyped markers with TUNA. Bioinformatics 24:435–437.

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. 2002. Truncated product method for combining *P*-values. Genet Epidemiol 22:170–185.