Published in final edited form as:

Hum Genet. 2013 July ; 132(7): 745–759. doi:10.1007/s00439-013-1285-4.

Recursive Organizer (ROR): an Analytic Framework for Sequence-Based Association Analysis

Lue Ping Zhao^{1,2,*} and Xin Huang¹

¹Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

²School of Public Health, University of Washington, Seattle, WA, USA

Abstract

The advent of next-generation sequencing technologies affords the ability to sequence thousands of subjects cost-effectively, and is revolutionizing the landscape of genetic research. With the evolving genotyping/sequencing technologies, it is not unrealistic to expect that we will soon obtain a pair of diploidic fully-phased genome sequences from each subject in the near future. Here, in light of this potential, we propose an analytic framework called, recursive organizer (ROR), which recursively groups sequence variants based upon sequence similarities and their empirical disease associations, into fewer and potentially more interpretable super sequence variants (SSV). As an illustration, we applied ROR to assess an association between *HLA-DRB1* and type 1 diabetes (T1D), discovering SSVs of *HLA-DRB1* with sequence data from the Wellcome Trust Case Control Consortium (WTCCC). Specifically, ROR reduces 36 observed unique *HLA-DRB1* sequences into 8 SSVs that empirically associate with T1D, a four-fold reduction of sequence complexity. Using *HLA-DRB1* data from Type 1 Diabetes Genetics Consortium (T1DGC) as cases and data from Fred Hutchinson Cancer Research Center as controls, we are able to validate associations of these SSVs with T1D. Further, SSVs consist of nine nucleotides, and each associates with its corresponding amino acids. Detailed examination of these selected amino acids reveals their potential functional roles in protein structures and possible implication to the mechanism of T1D.

Keywords

Gene; DNA sequence; HLA; Structural variations; Type 1 diabetes

Introduction

Recent developments of next-generation sequencing technologies is transforming the landscape of genetic association studies, allowing researchers to have a complete access to human genomes from normal and diseased subjects (Mardis 2008). These technologies, commercially available from, e.g., Illumina, 454 Life Science, Solid or Complete Genomics, are available to produce, on each subject, over 3.5 million single nucleotide polymorphisms (SNPs), over 0.5 million indels, and over one million copy number polymorphisms (CNPs), in addition to numerous structural polymorphisms, which are far greater than SNPs from typical SNP-based genomewide association studies (GWAS). Even more exciting is the

*Corresponding Author: lzhaol@fhcrc.org, 1100 Fairview Avenue North, P.O. Box 19024, Mailstop M2-B500, Seattle, WA 98109-1024, Phone: (206) 667-7280, Fax: (206) 667-2437.

Author Contributions

LPZ conceived the idea and developed the methodology, and XH developed and implemented the methodology, in addition to performing all of the numerical work.

potential development of phased sequencing technologies, to produce fully phased diploidic sequences from each subject (Yang, Chen et al. 2011; Peters, Kermani et al. 2012). When fully developed, these technologies will read out paternal and maternal genome sequences for each subject. Because of known phases, one should readily identify all possible point polymorphisms (such as SNPs) and structural polymorphisms (such as insertions/deletions, or rearrangements), which are created by meiotic mutations and recombinations throughout the evolution. Consequently, we would expect that everyone would have unique genome sequence pairs, with a possible exception of identical twins. At that time, the challenge was how to investigate genetic associations with human diseases, even if everyone on the planet were sequenced. With this ultimate analysis challenge in mind, we would most likely focus on more specific genome sequence features, such as structural changes on a chromosomal level, or regulatory sequence variations, or polymorphisms in functional genes.

Before carrying our thinking too far into the future, we propose studying Human Leukocyte Antigens (conventionally, HLA genes) as a model to investigate challenging analytic issues. HLA genes reside in the major histocompatibility complex (MHC), approximately 4 MB region on chromosome 6p (Supplementary Figure S1)(Stewart, Horton et al. 2004; Brown, Pierce et al. 2009). Despite its modest genome size, MHC has many typical genome sequence complexities, including point polymorphisms, structural polymorphisms, recombination hot spots, high linkage-disequilibrium (LD), and exceptionally high polymorphisms. Meanwhile, MHC is considered to be one of the most gene-rich regions in the genome with around 140 genes, many of which play essential roles in autoimmunity and innate immunity and associate with many complex diseases. To narrow our discussion further, we focus on HLA-DRB1 (also, HLA-DQB1 due to close proximity and high LD) in this manuscript because it is well-known for its association with several other autoimmune diseases(Thorsby and Lie 2005; Forabosco, Bouzigon et al. 2009). Probably, the best HLA-DRB1 and -DQB1 association is with type 1 diabetes (T1D)(Schober, Schernthaner et al. 1981; Horn, Bugawan et al. 1988; Sheehy, Scharf et al. 1989; Noble and Valdes 2011), including this haplotype *HLA-DRB1**15:01-*DQA1**01:02-*DQB1**06:02erlich 2008 (Erlich, Valdes et al. 2008). Also, HLA-DRB1 is among five HLA genes that are translated into hematopoietic stem cell transplantation (Petersdorf 2004; Malkki, Single et al. 2005). Most relevant to our discussion is that this gene is routinely sequenced in phases; phased sequences are coded as alleles of HLA-DRB1, and hence gene alleles can be converted into sequences. Hereafter, an HLA gene allele and sequence variant are used interchangeably. According to the current estimate, HLA-DRB1 and -DQB1 sequences are highly polymorphic with around 1200 and 179 alleles, respectively (<http://hla.alleles.org/>).

Facing this exceptionally high polymorphism, statistical geneticists have been actively developing innovative methods to establish genetic associations with diseases, in particular, HLA and T1D. For those extremely common HLA alleles or their common extended haplotypes, one could amass a sufficient number of study subjects for a meaningful statistical assessment, typically using a chi-square analysis of contingency table (Everitt 1986). To analyze other alleles or other genes, one needs to adjust for those disease-associated common alleles/haplotypes. If choosing matching cases and controls by those known HLA alleles/haplotypes, one would have dwindling sample sizes leading to loss of power. Further, when analyzing alleles with relatively low frequencies, one may have no power to assess their disease associations with those matched samples. Recently, Todd and his colleagues described an application of recursive partitioning method with the conditional logistic regression method to systematically group different alleles in assessing disease associations (Cordell and Clayton 2002; Nejentsev, Howson et al. 2007; Todd, Walker et al. 2007). While this approach has been endorsed in a recent T1D consortium (Rich, Akolkar et al. 2009), it groups alleles purely based upon their empirical associations, and hence may have limited capacity in controlling false positive errors like many data-driven methods.

Another promising approach is to explore sequence features of alleles, known as sequence feature variant type (SFVT) analysis (Karp, Marthandan et al. 2010; Thomson, Marthandan et al. 2010). However, SFVT method relies on the successful extractions of prominent sequence features, which can be high dimension and also dependent upon knowing protein structures of the target genes.

Following the motivations of the recursive portioning method and SFVT, a desirable method should center on grouping multiple alleles into some super alleles as well as retaining sequence integrity (i.e., the order of nucleotides). Here, we propose an analytic framework, known as recursive organizer (ROR), to recursively organize multiple sequence variants in the context of the disease association studies. ROR assumes sequence similarity index to identify those sequence pairs that are deemed similar to each other and hence are likely to have similar disease associations, while empirically tests the above assumption iteratively. Recursively, ROR is able to organize multiple alleles (sequences), regardless of their allelic frequencies, into fewer “super sequence variants” (SSV), while retaining their empirical association profiles. In this manuscript, we illustrate ROR through an application to the analysis of T1D association with *HLA-DRB1*.

Material and Methods

Data Sources

WTCCC—As one of most successful GWAS enterprises, WTCCC (Wellcome Trust Case Control Consortium) has completed scanning genomes for disease associations with seven complex diseases, including type 1 diabetes (T1D) together with a set of common controls (WTCCC 2007). Following this success, WTCCC has expanded the genotyping effort to include additional control samples. Of particular interest, HLA genotyping was performed on some of subjects (see below). Genetic data, together with phenotype data, are publicly available (<https://www.wtccc.org.uk/index.shtml>). WTCCC includes 2,000 T1D cases from UK National Health Service hospitals (WTCCC 2007). Additionally, it includes a total of 3,000 normal subjects from the birth cohort of 1958, as controls. Of all subjects, 1,639 cases and 1,140 controls have been genotyped on *HLA-DRB1* at high resolution, and are used in this study.

For the data access, we have applied and obtained genotype and phenotype data from the European Genome-Phenome Archive (<http://www.ebi.ac.uk/ega/page.php>), following the established procedure at European Bioinformatics Institute.

T1DGC—Being a research consortium, T1DGC (Type 1 Diabetes Genetics Consortium) has collected families, where one or more siblings are diagnosed with T1D. Sampling one case per affected family, we identified 3,760 cases diagnosed with T1D. All selected cases have been genotyped with *HLA-DRB1* at high resolution. These subjects are used as cases in the validation study.

The genetic data set is provided by the Type 1 Diabetes Genetics Consortium (<https://www.t1dgc.org>), a collaborative clinical study sponsored by multiple institutes of NIH and Juvenile Diabetes Research Foundation International (JDRF). We have followed the established procedure to apply and access this case data set from dbGAP (<http://www.ncbi.nlm.nih.gov/gap>).

HSCT—Being a premier cancer center on hematopoietic stem cell transplantation (HSCT), Fred Hutchinson Cancer Research Center has established a cohort of recipients and their donors, which is briefly described elsewhere (Chien, Zhang et al. 2012). For matching

purposes, most of donors are healthy and are genotyped for HLA genes. For this study, a total of 1,249 donors are included into the validation study as controls.

All participants involved in HSCT signed a consent form that allows uses of genetic data consistent with our methods here.

Sequence Data of HLA-DRB1

Genes in major histocompatibility complex (MHC) have been linked with T1D in earlier research (Todd, Bell et al. 1987; Cucca, Dudbridge et al. 2001; Nejentsev, Howson et al. 2007) and have been further confirmed by GWAS (Todd, Walker et al. 2007; Bradfield, Qu et al. 2011). Within MHC, *HLA-DRB1*, together *HLA-DQB1*, is the most important genes associated with T1D. Through decades of research on their associations with T1D, two common alleles, DR3 and DR4, have unambiguously been found to associate with T1D (Schober, Scherthaner et al. 1981; Horn, Bugawan et al. 1988; Sheehy, Scharf et al. 1989; Noble and Valdes 2011), although further delineation of T1D associations with other less common alleles has been challenging. In light of what has been discovered, it is important to dissect T1D associations with less common alleles of *HLA-DRB1*. Note that due to the long history of researching HLA genes, common alleles, as opposed to rare alleles, have been formally defined elsewhere (Cano, Klitz et al. 2007). In this manuscript, we use uncommon alleles to refer those alleles that have fewer than 5 copies in the empirical data, even though the definitions of common and rare alleles in GWAS literature are slightly different.

For all *HLA-DRB1* alleles (4 digits) reported in the above three sources, we extract corresponding sequences. In WTCCC, there are 36 distinct 4-digit resolution alleles, of which 9 are uncommon alleles (01:04, 03:02, 08:03, 08:06, 12:02, 13:05, 13:10, 14:04, 16:02) (Table 1). To obtain the nucleotide sequences of these 36 distinct 4-digit resolution alleles, we first extracted their 6-digit resolution nucleotide sequences from the IMGT/HLA database (Release 3.5.0, 14th July 2011, www.ebi.ac.uk/imgt/hla/). Typically, there are several possible 6-digit resolution sequences for each specific 4-digit HLA allele. We then extend the coding schema to capture several possible nucleotide combinations at each nucleotide position via the IUPAC-IUB symbols for nucleotide nomenclature (Cornish-Bowden 1985). By this coding strategy, our procedure allows one to “impute” the entire allelic sequences, including exon 2 and 3, even though the original typing is restricted within exon 2 (Nejentsev, Howson et al. 2007; Mychaleckyj, Noble et al. 2010). The sequence variant *HLA-DRB1**01:01 is used as a reference in the association analysis, for easy interpretation when comparing with existing literature. Similarities between all sequence pairs are measured as a percentage of nucleotide identities base-by-base (to be defined below), and are used to construct the heat map and a coalescent tree (see Supplementary Figure S2). Alleles in the heat map (Figure S2a) are ordered by their empirical associations with SSVs (Table 1). It shows that those alleles within the same SSV tend to have similar sequences but similar sequences do not necessary have the same empirical associations. The companion coalescent tree (Figure S2b), on the other hand, shows the evolutionary relationship among these 36 classic HLA alleles, providing an insight into sequence relationship beyond their nomenclature relationship.

The ROR Procedure

The ROR algorithm takes phenotype and diploidic sequences from a case-control association study as input and produces SSVs after merging similar sequence variants with comparable disease associations. The recursive procedure starts with a panel of distinct sequence variants in the study population, dynamically removes nucleotide(s) that differentiate between sequence variants, and merges corresponding sequence variants into

SSVs if such merging does not alter disease association (Figure 1). To be specific, we assume that the genetic penetrance model follows the logistic regression model. Suppose n case-control subjects, with cases denoted by $d_i = 1$ and controls denoted by $d_i = 0$ where the subscript $i=1,2,\dots,n$ denotes subjects, are sequenced with L nucleotides in length, resulting in a pair of phased sequences $(\dot{h}_i, \ddot{h}_i)$, where $\dot{h}_i = \dot{h}_{i1} \dot{h}_{i2} \dots \dot{h}_{iL}$ and $\ddot{h}_i = \ddot{h}_{i1} \ddot{h}_{i2} \dots \ddot{h}_{iL}$ and a_{ij} 's are individual nucleotides. Again, this notation formalizes the equivalence between a haplotype of multiple nucleotides and a phased sequence. The ROR procedure consists of the following steps:

Step 1 is to evaluate similarities between all possible haplotypes. Suppose that one has a set of haplotypes $(\underline{h}_1, \underline{h}_2, \dots, \underline{h}_m)$, and their frequencies are denoted as $(f_1, f_2, \dots, f_m)$ where haplotype frequencies are greater than zero. In the context of *HLA-DRB1*, haplotypes and haplotype frequencies are known as alleles and allelic frequencies, respectively. Now, to evaluate pairwise similarities, we use the following metric as

$$S_{j,k} = \Phi(\underline{h}_j, \underline{h}_k | w, f_j, f_k) \quad (1)$$

Where $j > k$, $\Phi(\underline{h}_j, \underline{h}_k | w, f_j, f_k)$ is a symmetric kernel function quantifying certain desired features in the assessment of sequence similarity, and $w = (w_1, w_2, \dots, w_L)$ is the weight vector with positive weights and unit sum. How to choose such kernel function with appropriate weights will be considered in the following section.

Step 2 is to merge haplotypes that are deemed to be most “similar”. In the current application, the kernel function is to compare sequence identity nucleotide-by-nucleotide. Hence, two haplotypes deemed to be highly similar to each other are likely to differ at only few nucleotides. By removing nucleotides that differentiate two haplotypes, one would merge corresponding haplotypes.

Step 3 is to assess if the proposed merge alters the genetic associations with the phenotype, the statistical information which is quantified by the likelihood. In other words, merging should NOT alter the likelihood, if those removed nucleotides are not playing essential roles in the empirical associations, or if disease associations are retained within reduced sequences. Otherwise, the likelihood will be “significantly altered” with the merged haplotypes. To evaluate the association under the likelihood, we use the following logistic regression model for modeling the penetrance of haplotypes to disease phenotype, e.g.,

$$\Pr(d_i=1|\dot{h}_i, \ddot{h}_i) = 1 / (1 + \exp\{-\alpha - \beta' [I(\dot{h}_i) + I(\ddot{h}_i)]\}) \quad (2)$$

where $[I(\dot{h}_i) + I(\ddot{h}_i)]$ is a vector of indicator sums for 0, 1 or 2 copies of all possible haplotypes, and (α, β) are regression coefficients to be estimated. The above model can be adapted to model genotypic associations, dominant or recessive associations. Under the above model, one can compute the corresponding log likelihood function $\ell(\alpha, \beta)$. Now if by removing a few pertinent nucleotides, one can compute a reduced log likelihood function with fewer haplotypes, which is denoted as $\ell(\alpha^r, \beta^r)$. To evaluate the empirical associations with those removed nucleotides or with those merged haplotypes, one can compute the following log-likelihood ratio statistic:

$$\chi^2_{df} = 2[\ell(\alpha, \beta) - \ell(\alpha^r, \beta^r)], \quad (3)$$

where the above statistic has a chi-square distribution with the degree of freedom being the number of merged haplotypes, if the merging does not alter the empirical association. This

chi-square statistic can thus be used to compute the significance p-value, based upon which the statistical inference can be made. Given a pre-set threshold, say 0.05, one would proceed to merge haplotypes, if the associated p-value is greater than the threshold. Otherwise, one may skip the proposed merging and examine the possibility of merging the next group of similar sequences.

Step 4 is to repeat steps 1–3, until exhausting all possible merges based upon p-value assessment. The ROR may terminate the procedure, either when all possible pairs are exhausted, or no additional nucleotides can be removed. In each ROR step, starting from the most similar haplotype pair ($r=1$), the log-likelihood ratio test rejects the null hypothesis if the p-value is smaller than the pre-set threshold value α . Now if the first chosen haplotype pair is not merged, the ROR procedure begins to assess the significance of empirical association for the second most similar haplotype pair ($r=2$). To adjust for multiple comparisons, the threshold value is adjusted to α/r for assessing the significance.

After removing nucleotides that are not likely causal, ROR produces a set of haplotypes, much fewer than the original set, which are referred to as SSVs. In comparison with the reference SSV, those SSVs with positive associations are deemed as risky haplotypes, while those with negative associations are deemed as protective haplotypes.

Sequence Similarities

As noted above, a metric for measuring similarity between sequence pairs is an important component of ROR, reflecting the study assumption that “similar sequences” are likely to have similar functions and hence similar empirical associations. A relatively simple and yet, probably, the most commonly used metric bases on identity at every nucleotide, assuming full alignment among sequences. Specifically, the similarity metric used in equation (1) may be written as

$$\Phi(\underline{h}_j, \underline{h}_k | w, f_j, f_k) = W(f_j, f_k) \sum_{l=1}^L w_l I(a_{jl} = a_{jk}), \quad (4)$$

where $W(f_j, f_k)$ is a weighted function of two haplotype frequencies, e.g., $W(f_j, f_k) = f_j f_k$ to over-weight those common haplotypes for merging purposes, the indicator $I(a_{jl} = a_{jk})$ equals one if the equality is true, and w_l is the nucleotide-specific weight with the unit sum

$\sum_{l=1}^L w_l = 1$. The nucleotide-specific weight w_l can be chosen to reflect the functional significance of the corresponding nucleotide. For example, one possible choice is to assign the non-synonymous nucleotides to have a twice the weight of the synonymous nucleotides, synonymous nucleotides to have a twice the weight of those nucleotides in introns, and weights in the flanking regions are monotonically decreasing as corresponding nucleotides are distant from the gene. If the non-synonymous is favored with the larger weight, ROR would tend to group those alleles with the same amino acids.

There are alternative similarity metrics to the nucleotide-identity measurement in defining $\Phi(\underline{h}_j, \underline{h}_k | w, f_j, f_k)$. For example, “counting measure” quantifies the similarity by the proportion of SNPs at which two haplotypes are the same, or “length measure” quantifies the length of the longest continuous interval of matching alleles shared between two haplotypes. As discussed by Tzeng et al. (Tzeng, Devlin et al. 2003), the length measure has advantages in capturing partial sharing due to recombination in the ancestral haplotype, but it is not robust to genotyping errors, missing data, and recent mutations. The counting measure, meanwhile, is robust in the above scenarios, and it has convenient statistical

features. Another popular metric bases on genealogical or cladistic distance (Tachmazidou, Verzilli et al. 2007; Bansal, Libiger et al. 2010), which requires strong phylogenetic assumptions (Bansal, Libiger et al. 2010). All the above similarity metrics are appropriate depending on the scientific perspective on what is the essential feature for the sequence similarity. In the presence of repeats or copy number variations, one may seek other metrics for measuring the difference between sequence pairs (Schork, Wessel et al. 2008). For example, Varre et al. quantified the similarity of two sequences in terms of segment-based events (Varre, Delahaye et al. 1999). By calculating the minimal number of segment operations needed to transform one sequence to the other, they defined a family of similarity measures called transformation distances. The transformation distance is essentially the minimum description length among all possible scripts that build two sequences. In the context of HLA genes for which their tertiary structures have been established, probably more meaningful similarity between sequences should be their similarity of “tertiary structures” or of “critical structures” formed by a subset of amino acids.

Permutation and False Positive Error Rate

While intuitive, ROR involves recursive computations with dynamically merging sequence variants based upon both sequence similarities and their association statistics. Due to the sequential nature, it is difficult, if not impossible, to derive explicit expressions that allow estimation of false positive error rates. Yet, without appropriately quantifying false positive error rates, ROR would degenerate into an ad hoc method with unknown false positive error rates. Towards this goal, we propose a permutation approach to quantify false positive error rate. Briefly, permuting disease phenotypes across all subjects, one has a realization of sequence-phenotype data under the null hypothesis. Repeating ROR on each permuted data set, say 1,000 times, would allow one to estimate false positive error rate, which is customarily controlled at 1% or 5%.

Bootstrap Analysis and Power Estimates

Power is complementary to false negative error rate, i.e., it is the probability to discover the true association signals. Power evaluation is typically required for designing studies. While the opportunity to derive an explicit formula for computing power of ROR is limited, we suggest using a re-sampling technique to evaluate powers. From an existing data set, one can sample, with replacement, individual sequence and phenotype data with the fixed sample size of the target size. On the sample data set, one performs the ROR, and evaluates the frequencies of positive discoveries from, say 1,000 replicates. Estimated frequencies approximate power estimates at given odds ratios.

Recursive Organizing Diagram (ROD)

To assist the interpretation, we developed a visualization tool, referred to as recursive organizing diagram (ROD), to document the organizing process by ROR. ROD is represented by a color circle, which records the organizing process and is analogous to an unrooted tree. A ROD shows the sequence variant names as leaves, the organizing step numbers as the nodes, and the number of removed nucleotides as the length of branches between two nodes (ROR steps). The shedding colors of a ROD correspond to direction of association with the density associated with the association magnitude, and its branch colors indicate significance levels (p-values). The center of the ROD is the plot of log likelihood increments for every ROR step.

Software

Software, ROR, implementing the proposed approaches is available freely on Quantitative Genetic Epidemiology group website (<http://qge.fhcr.org>).

Results

Type 1 Diabetes

As noted above, HLA genes have been found to associate with T1D in traditional HLA genetic studies and also in GWAS, and one of them is *HLA-DRB1*, together with *HLA-DQB1*. Further, fine-mapping efforts have pinpointed two common alleles DR3 and DR4 that are primary contributors to the T1D associations (Schober, Scherthaner et al. 1981; Horn, Bugawan et al. 1988; Sheehy, Scharf et al. 1989; Noble and Valdes 2011). However, investigating T1D associations with other *HLA-DRB1* alleles has been a challenge because of high polymorphisms and thus relatively modest allelic frequencies for most common alleles (Little and Parham 1999; Horton, Wilming et al. 2004). All observed *HLA-DRB1* alleles are listed in Table 1, including a total of 36 alleles in this data set. Further, investigating T1D associations with other HLA is hampered by extended LD between genes within *MHC* region (Miretti, Walsh et al. 2005; de Bakker, McVean et al. 2006; Santiago, Li et al. 2009).

Discovery Analysis

For the discovery analysis, we used the significance level of 1% to recursively organize 36 *HLA-DRB1* alleles (Figure 2). As an example, ROR in its first step merges *HLA-DRB1**16:01 and *16:02, then *12:01 and 12:02 in the second step, and eventually, *04:05 merged into a group (*04:01,*04:02, *04:04) in the 27th step. Log likelihood ratio statistics, from ROR step 1 to 27, are shown in the center of Figure 2. At the termination of ROR, a total of 8 distinct SSVs are resulted from the procedure, and are labeled as *SSV*#01 to *SSV*#08. Table 2 lists those SSVs and their corresponding association statistics from the discovery analysis. With the reference #01 (= *HLA-DRB1**01:01), alleles #02~#05 are protective, whereas #07 and #08 are risk alleles. *SSV*#02 and #03 are highly protective variants with the reduction of risks on the order of 20 to 50 times than the risk with individuals of carrying *SSV*#01. Even for *SSV*#04 and #05, they confirm a reduction of risk by 4 times as compared to the reference allele. In contrast, #07 and #08 have elevated risk of T1D by 1.5 to 2 times, in comparison with the reference allele. To relate these results back to actual *HLA-DRB1* alleles and known disease associations, we have listed corresponding alleles in the far right column of Table 1. Interestingly, all reported associations in the literature are consistent with identified risk profiles here. Additionally, the grouping by SSVs suggests that other alleles, which to our knowledge have not been reported yet, may share the same risks with those classified into the same SSVs.

Investigating robustness of discovered SSVs, we performed a bootstrap analysis, i.e., re-sampling 1,000 times with replacement from the case-control study with same and half sample sizes, and performing the same association analysis to estimate powers of discovering corresponding SSVs. The *SSV*#06, which is estimated to have an odds ratio around 1 (null hypothesis), confirms the power of 0.01, which is equivalent to the type I error rate established for the association analysis. For protective allele *SSV*#02 to #05, powers of detecting their associations are nearly 100%, and remain to be so, even with half of the original sample size (Figure S3). For risk alleles *SSV*#07 and #08, powers are around 60% or 97%, respectively.

To complement to the bootstrap-based power analysis, we also evaluate the false positive error rate by permutation. To proceed, we performed ROR on permuted data sets, by randomly permuting disease phenotypes across all subjects and generating a realization of case-control data. Hence, any discovery by ROR is false. Repeating this permutation 1000 times, we estimate the false positive error rate, i.e., the type I error rate, as the fraction of

false discoveries in all permutations. The estimate is slightly below 0.01 (not shown), which supports the validity of ROR.

Validation Analysis

For the validation analysis, we used 3,760 cases from T1DGC and 1,249 controls from HSCT. Table 1 lists their allelic frequencies among cases and controls (the right hand panel of Table 1). Allelic frequencies among cases and among controls in the validation cohort are largely comparable with those in the discovery cohort. Such consistencies are expected, since all subjects in these two cohorts are Caucasoid. Additionally, these consistencies are supportive of genotype data quality. Now on those discovered SSV#01-#08 (left hand panel in Table 2), we evaluated the same SSVs on the validation cohort. Estimated allelic frequencies, odds ratios, confidence intervals and corresponding p-values are presented in the right hand panel of Table 2. It appears that estimated allelic frequencies of these SSVs are comparable between two cohorts, despite minor differences. More importantly, when examining estimated association parameters, all estimated odds ratios in the validation analysis are in the same direction as those in the discovery analysis. With SSV#01 as the reference allele, SSV#06 remains null. SSV#02 to #05 remain significantly protective. On the other hand, SSV#07 and #08 are highly significant as risk alleles with smaller p-values, probably because of larger sample size and greater odds ratios. Overall, the validation analysis supports associations of T1D with all six discovered SSVs with high confidence.

Comparison with an Amino Acid-Based Procedure

Recently, Raychaudhuri and colleagues described a stepwise regression analysis to select significantly associated amino acids, and to evaluate their haplotypic associations with haplotypes of selected amino acids (Raychaudhuri, Sandor et al. 2012). Following the same principle of typical stepwise regression technique (Cordell and Clayton 2002), this procedure starts from predicting HLA alleles from SNPs, i.e., phased sequences for target genes, to converting those sequences into sequences of amino acids, and to progressively identifying significantly associated amino acids via a forward-stepwise fashion. Upon selecting those amino acids, the procedure then evaluates disease associations with haplotypes of those selected amino acids, for result interpretation. When applying this innovative procedure to seropositive rheumatoid arthritis, they identified five amino acids in HLA proteins that explain most of the disease association and produced grouping of classical HLA alleles corresponding to combinations of amino acids (Table 1 in their paper).

Raychaudhuri's approach shares the analytic objective with the ROR, but also has subtle differences. Indeed, both approaches aim to establish a minimum set of "super variants", in the form of haplotypes of amino acids or SNPs, and these super variants are deemed to explain most disease associations. However, the former approach chooses to use amino acids in sequences as a way to overcome excessive polymorphisms in HLA sequence data via the regression analysis, and then to evaluate their haplotypic associations only with marginally selected amino acids. In contrast, ROR deals with excessive polymorphisms in HLA sequences data directly, through recursively organizing all sequences via both sequence similarities and empirical associations. Consequently, ROR has better control on type I errors, while achieving a more parsimonious set of super variants that associate with disease phenotype.

To gain an insight into their empirical performance, we compare these two approaches on T1D association with *HLA-DRB1*, using WTCCC data set. To maximize the consistency, we use the fully phased sequence data, derived from *HLA-DRB1* alleles at high resolution, and perform association analyses using both approaches. Raychaudhuri's approach retain 4 amino acids (position 67, 74, 86, and 96), and haplotypes of these 4 amino acids result in 24

variants, each of which associates with a set of *HLA-DRB1* alleles (Table 3). Conceptually, this approach has reduced the polymorphisms from 36 unique alleles down to 24 alleles. Many of these 24 variants have varying associations with T1D, evidenced by odds ratios, confidence intervals, and p-values. Interestingly, association results are largely consistent to those obtained by ROR. For comparison, SSVs obtained by ROR are also listed in Table 3. SSV#1-3 are uniquely concordant between two approaches. SSV#04 correspond to four haplotype variants obtained by Raychaudhuri's approach. Further, the allelic association with SSV#04 is largely consistent with those individual associations obtained by Raychaudhuri's approach. Similar results hold up for other SSVs. In this example, ROR achieves even further reduction of polymorphisms from 24 alleles down to 8 alleles, without compromising empirical disease associations.

A Simulation Study

To assess the validity of ROR as well as to compare ROR with the amino acid-based stepwise selection approach by Raychaudhuri et al, we conducted a simulation study, following the T1D data from WTCCC. Specifically, we take allelic frequencies of *HLA-DRB1* from the WTCCC controls as the target population. From this population of alleles, we randomly sample two alleles from this population to form a diploidic individual and to compute the associated penetrance probability based upon the logistic regression model with the intercept of -5 (disease prevalence of $\sim 0.67\%$ for the reference group). We consider two simulation scenarios: First, amino acid-centric associations; an amino acid locus is randomly chosen as a causal locus as long as provided that one of alternative amino acids to the most common one has the allelic frequency greater than 5%. This alternative amino acid is set as the causal amino acid, and the corresponding penetrance is computed with a predetermined odds ratio. Second, haplotype-centric associations; a haplotype corresponding to amino acid at positions 67, 74, 86 and 96 is randomly chosen as the causal variants as long as its frequency is greater than 5%, and the corresponding penetrance is computed with predetermined odds ratios. Given the penetrance probability, we simulate a binary phenotype as a Bernoulli process: the phenotype equals to one (case) if a uniform random number exceeds the penetrance probability, otherwise equals to zero (control). Repeating this phenotype simulation process, we retain the first 1000 phenotypes of zero as controls, and the first 1000 phenotypes of one as cases. On each sample of cases and controls, we performed both ROR and Raychaudhuri's approaches, obtaining a test statistic for this single realization. Repeating the same simulation and analysis process 500 times, we estimated the percentage of rejecting the null hypothesis, under a specific alternative scenario of particular odds ratio. We set the odds ratio ranges from 1 to 1.5 under the first scenario, and from 1 to 2.5 under the second scenario.

Odds ratios of 1 under either scenario 1 or 2 correspond to the null hypothesis of no association. For both analyses, we set the overall type I error rate at 5%. Estimated error rates are approximately 0.04, 0.03 under scenario 1 and 0.03, 0.01 under scenario 2, for ROR and Raychaudhuri's analyses, respectively. While both approaches are somewhat conservative, Raychaudhuri's approach is slightly more conservative due to Bonferroni's correction. ROR is conservative, largely due to intrinsic correction of more than one test in some ROR steps.

Figure 3 shows the power curves by both approaches under two different scenarios. Under scenario 1 (left panel of Figure 3), the power of ROR is comparable to Raychaudhuri's approach of the amino acid-based stepwise selection procedure. This result is expected, since the simulated disease causal association is directly with individual amino acids. Hence, haplotype-based ROR has limited power gain over the amino acid-based approach. Under the second scenario, however, ROR is consistently more powerful than the approach of the amino acid-based stepwise selection procedure. This power differentiation is due to the fact

that the amino acid-based approach extracts most of the marginal associations, without benefiting from their haplotypic associations that are captured by ROR.

Discussion

We have described an analytic framework, ROR, for organizing polymorphic phased sequences into fewer and more meaningful SSVs. Just as with the reduction of DNA sequence into amino acid sequence, for those in coding regions, the reduction by ROR represents yet another level of information simplification, enabling us to focus on key polymorphisms that are directly associated with disease phenotypes. As a framework, ROR is applicable to phased sequence data, from specific candidate genes or regions, to case-control studies with binary phenotypes.

Applying ROR to analyze *HLA-DRB1* and its association with T1D in a case-control study as an illustration, we started with *HLA-DRB1* sequence data of over 800 nucleotides and 36 sequence variants, and performed 27 ROR steps, reducing its polymorphisms to 8 SSVs with 9 SNPs; located at positions 17t, 23t, 24t, 37t, 60f, 74s, 75t, 98f, 114t, where the number corresponds to the position of the amino acid and the letter indicates the position of the nucleotide in the amino acid (f=first, s=second and t=third). In particular, position 74 was actually reported earlier as an important amino acid with strong association with autoimmune diseases including T1D (Ban, Davies et al. 2004; Menconi, Osman et al. 2010), a positive control for ROR. Now SSV#02 to #05 are all found to be protective, in comparison with the reference SSV#01, while SSV#07 and #08 are found to increase risk. Interestingly, obtained groupings of these disease-associated alleles are consistent with known associations in the literature (Table 1) (Cucca, Dudbridge et al. 2001; Koeleman, Lie et al. 2004; Nejentsev, Howson et al. 2007).

Beyond being consistent with the literature, SSV groupings potentially offer additional insights into other DR alleles that have relatively low frequencies. For example, via ROR analysis, it is found that *HLA-DRB1**07:01 and *09:01 have been grouped with *HLA-DRB1**04:03 and *04:07 in SSV#04. Based on the ROR grouping, one would infer that both *HLA-DRB1**07:01 and *09:01 alleles are protective alleles, associations for which have yet to be reported in the literature. Indeed, with a relatively high allelic frequencies among controls (~18% and 13% in discovery and validation cohort, respectively), its protective association may be thought to validated. In contrast, the allelic frequency for *HLA-DRB1**09:01 is only 1.8% in discovery controls and 1.0% in validation controls, in comparison with their respective frequencies of 1.3% and 1.8% among cases, would not immediately be supportive of any protective associations. However, this allele is grouped to SSV#04, largely because of its sequence similarity. Intuitively, for relatively uncommon alleles, one would tend to rely on sequence similarity to judge functional associations. Nevertheless, it is interesting to note a recent report that both *HLA-DRB1**07:01 and *09:01 associate with autoantibodies to islet antigen-2 (Williams, Aitken et al. 2008). Another example is *HLA-DRB1**03:02, which is observed only once among both discovery and validation cases. This allele is absent among controls. Again, by the sequence similarity, ROR grouped this allele with *HLA-DRB1**03:01, and infers that this allele is a risk allele. Of course, we need to be cautious about conclusions on these uncommon variants, since SSV grouping relies on the implicit assumption that similar sequences likely share similar function. For example, *HLA-DRB1**01:02 seems to be more common among cases (0.92%) than among controls (0.79%), but it is merged into SSV#05 as a protective allele with many other alleles. Interestingly, in the validation data set, *HLA-DRB1**01:02 becomes less common among cases (0.57%) than among controls (1.4%), which likely indicates the implication of its protective effect as predicted by ROR.

One key advantage to using SSVs is that they are linked with amino acid sequences and thus protein structures. For example, 8 SSVs identified here correspond to a total of fifteen possible codon sequences (see Supplementary Table S1). To allow some unambiguity in their sequences, we use letters (K, R, C, M, B) to indicate variable nucleotides at those loci. Collectively, codon sequences are converted to amino acid sequences (reference sequence “FRVSYAVKL”). Note that within SSV#07 and #08, their codon sequences vary, but amino acid sequences are the same. For SSV#05, corresponding amino acid sequences appear to be more variable at position 37 and 60. The SSV#04 appears to be less heterogeneous, with variable amino acids at position 37, 60, 74, with the 98th amino acid being E. Further, these SSVs are naturally linked up with their corresponding protein structures (see Supplementary Figure S4), which facilitate functional validations at structural levels.

A puzzling and yet interesting observation is that amino acid sequences associated with SSV#01 and #02 are the exactly same, and yet SSV#02 is highly protective. Based on empirical data from analyzing HLA-DRB1, one would conclude that the protective association with SSV#02 must be through other genes that are in high LD, or through regulations of non-coding genetic variations. Indeed, the classic allele *HLA-DRB1**15:01 in SSV#02 is known on a haplotype of *HLA-DRB1**15:01-*DQA1**01:02-*DQB1**06:02 (Erich et al. 2008). *DQB1**06:02 is known to be protective of T1D. So this unusual association may be mediated through LD between DR15-DQ06.

While being intrigued with new insights here, it is important to note a key limitation that needs to be accounted when evaluating current results. Due to utilizing existing HLA data from WTCCC and T1DGC, there is an inherent ambiguity from the SSOP typing method, i.e., nucleotide polymorphisms in exon 3 are not fully accounted for (see official 510(k) filing on the SSOP method by DYNAL, recorded by Food Drug Administration). Thus, the inference of exon 3 nucleotide sequence is not reliable. Hence, identified SNPs in exon 3 need to be interpreted with caution, even though this limitation does not negatively impact on validity of demonstrating the ROR methodology, the primary intent here.

When performing association analysis, we have chosen consistently the *HLA-DRB1**01:01 as the reference SSV. While this choice is reasonable for analyzing this gene given the need to compare results with the long-standing literature, a general approach would simply use the most common SSV as a reference, treating all variants equally. If this strategy were used on the current data set, results would be largely comparable (not shown).

Because of its hybrid approach of utilizing both sequence similarity and empirical association, ROR has several properties that are worthy elaborations. ROR is designed to focus on polymorphisms of sequences. Some of the known sequence polymorphisms may include single nucleotide polymorphisms, insertions/deletions, duplications and short-tandem repeats. To accommodate different types of polymorphisms, ROR needs to use appropriate metrics of measuring sequence similarity. For example, rather than percentage of nucleotide identities as used here, one can use numbers of mutations for one sequence polymorphism to “mutate” into another sequence polymorphism, similar to the kinship coefficient (Zhao and LeMarchand 1992). With this modification, ROR can readily assess disease associations with complex structures.

Another feature of ROR is its robustness to the presence of sequencing errors in the association analysis. Due to the nature of current short-read sequencing technologies, sequencing errors are inevitable, and are estimated around 0.5% or less. Sequencing errors are typically random and may be assumed to be independent. Hence, the presence of a sequencing error typically would produce “rare sequence variant”, often singletons. When such sequence variants are analyzed by ROR, many of them would be grouped with

sequences based upon their sequence similarities. Consequently, ROR is able to robustly correlate the target sequence variant with disease phenotype, without sacrificing statistical power.

As noted above, ROR is applicable to phased diploidic sequences, and thus does not need to infer phases, unlike other haplotype-based methods. For this reason, ROR does not need the assumption of the Hardy-Weinberg equilibrium (HWE) that is typically required for haplotype-based methods (Li, Khalid et al. 2003). In this respect, ROR is more robust when being applied to genes, such as MHC, where HWE is typically violated among cases.

Finally, ROR is suitable for dealing with uncommon variants (some of which are also referred to as rare variants in GWAS literature) in disease associations. Uncommon variants are frequently observed in disease association studies, when sequencing technologies are utilized (Cirulli and Goldstein 2010). To assess disease associations with such uncommon variants, one has no choice but to rely on known gene annotations, sequence variations, and sequence similarity to known sequence variants. Towards this goal, ROR compares these uncommon variants with more common sequences and merges them into SSVs whenever appropriate. If appropriately applied, ROR maybe one of the more effective strategies to make inference on uncommon variants, which is consistent with the idea of merging rare variants (Liu and Leal 2010). In this illustrative example of ROR, many uncommon alleles are merged into SSV#05. Indeed, this feature of ROR distinguishes itself from the conventional analysis of contingency table analysis that assesses disease associations with multiple alleles (Wessel and Schork 2006). In other words, ROR intelligently groups those uncommon alleles with common, and then assesses allelic associations with the disease outcome. Further, ROR allows one to adjust for confounding factors via the logistic regression models.

While appreciating the grouping feature of ROR, one could ask how ROR is different from a contingency table analysis of DRB1 with two-digit resolution, which may be thought of as grouping alleles that share the same critical amino acid sequences. Actually, the key difference is that the assumption required by ROR is weaker than the contingency table analysis. Specifically, the analysis with only two-digit resolution assumes that allelic variations with the same first two digits, i.e., same amino acid sequences, must have the same empirical associations, since nucleotide polymorphisms coded by third and higher digits are non-functional. The best illustrating example for this difference is the discovery of yin-yang effects of HLA-DRB1*04, i.e., HLA-DRB1*04:03 and *04:07 are protective and yet HLA-DRB1*04:01, *04:02, *04:04 and *04:05 are risky. Using a naïve two-digit analysis, one would miss this important discovery. In contrast, ROR has successfully identified this diametrically opposing effect between SSV#04 and #08.

While the hybrid nature of ROR, to the best of our knowledge, may be novel, the stepwise or recursive nature of the procedure has been used in other methods reported in the genetics and statistics literature. While reviewing this literature is not the objective here, we want to highlight several key references to two classes of methods. One class of methods, arising from population genetics, is to evaluate sequence similarities/differences, and to merge relevant sequences by their evolutionary lineages (Durrant, Zondervan et al. 2004; Tachmazidou, Verzilli et al. 2007). In contrast, another class of methods is based more on empirical associations, best represented by CART (Segal 1988) or by recursive partitioning (Nejentsev, Howson et al. 2007). These methods systematically evaluate empirical associations with all sequence variants, and will merge them if empirically acceptable. Taking advantage of both genetic sequence features and empirical associations, the method SFVT is to assess empirical disease associations with an array of biologically meaningful sequence features that are extracted from sequence data (Karp et al. 2010; Thomson et al.

2010). The success of carrying out SFVT hinges on the availability of well-established and documented sequence features. Sharing the same idea with SFVT, ROR also borrows information from sequence similarity information and empirical associations, except that ROR does not require any detailed gene-specific information, such as amino acids in protein structures, other than conventional gene annotations. Probably, the closest method to ROR is one that bases the analysis of disease association on haplotype evolution: first building a coalescent tree for haplotypes, and then assessing association or reverse (Buntjer, Sorensen et al. 2005). However, what distinguishes them is that ROR is making sequence similarity and empirical associations at every step, and then dynamically merges corresponding sequence variants recursively.

ROR shares the same analytic objective as a recently described forward stepwise procedure by Raychaudhuri et al., and they are complementary. Besides technical differences between these two approaches, ROR is designed specifically to discover a parsimonious set of sequences that associate with disease phenotype. Consequently, ROR can achieve much greater reduction of sequence polymorphisms associated with a specific disease phenotype. On the other hand, Raychaudhuri's approach is designed to look for "associated amino acids". Through conditional analyses, this approach can be more effective to pinpoint amino acids that have strongest empirical associations with disease phenotypes, even though the inferred associations are not necessarily causal.

There are several limitations. First of all, ROR is developed for phased sequences, prohibiting a direct application to GWAS where SNPs or nucleotides from NGS are unphased. Given the routine availability of unphased SNPs from GWAS, it is important to extend ROR to incorporate uncertainty of phases and to assess empirical associations via haplotype-based methods (Zhao, Li et al. 2003). It is important to note that ROR is fundamentally a haplotype-based method, when phases are known. Secondly, the current ROR is restricted to the binary phenotypes arising from case-control GWAS. In reality, GWAS is increasingly used to discover genes associated with quantitative phenotypes or with censored phenotypes. To broaden the appeal of ROR, it is necessary to introduce generalized linear model as a model to assess genetic associations so that ROR would be applicable to other types of phenotypes. Thirdly, we used T1D association with HLA-DRB1 as an illustrative application of ROR, but it is known to be in high LD with HLA-DQB1. Further, other class HLA genes, including HLA-A, -B, -C and -DP, may play a role in T1D, and, potentially, other minor antigens within MHC may also be important for T1D. To do a full justice for T1D research, it is important for us to drill deeper into their genetic associations, shedding new insights into genetic mechanisms of T1D associations with MHC.

In summary, this manuscript has described ROR for analyzing fully phased diploidic sequences. Conceptually, ROR utilizes both sequence similarity and empirical association to guide the recursive organization of multiple sequences into relatively fewer sequences. Statistically, ROR is shown to have desirable properties with respect to retaining analytic power and false positive error rates. Practically, ROR is applied to examine the T1D association with HLA-DRB1, one of most important and polymorphic genes in the human genome, results from which are validated and encouraging. In addition to the several future developments identified above, developing ROR could show that such a method is scalable to analyze phased whole genome sequences in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Authors would like to thank two anonymous reviewers for their helpful comments that significantly improve the presentation of the method and its application. This work is supported in part by NIH grant R01 MH084621.

References

- Ban Y, Davies TF, et al. Arginine at position 74 of the HLA-DR beta1 chain is associated with Graves' disease. *Genes Immun.* 2004; 5(3):203–208. [PubMed: 15029234]
- Bansal V, Libiger O, et al. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet.* 2010; 11(11):773–785. [PubMed: 20940738]
- Bradfield JP, Qu HQ, et al. A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet.* 2011; 7(9):e1002293. [PubMed: 21980299]
- Brown WM, Pierce J, et al. Overview of the MHC fine mapping data. *Diabetes Obes Metab.* 2009; 11(Suppl 1):2–7. [PubMed: 19143809]
- Buntjer JB, Sorensen AP, et al. Haplotype diversity: the link between statistical and biological association. *Trends Plant Sci.* 2005; 10(10):466–471. [PubMed: 16154381]
- Cano P, Klitz W, et al. Common and well-documented HLA alleles: report of the Ad-Hoc committee of the American society for histocompatibility and immunogenetics. *Hum Immunol.* 2007; 68(5):392–417. [PubMed: 17462507]
- Chien JW, Zhang XC, et al. Evaluation of published single nucleotide polymorphisms associated with acute graft versus host disease. *Blood.* 2012
- Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet.* 2010; 11(6):415–425. [PubMed: 20479773]
- Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet.* 2002; 70(1):124–141. [PubMed: 11719900]
- Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.* 1985; 13(9):3021–3030. [PubMed: 2582368]
- Cucca F, Dudbridge F, et al. The HLA-DPB1--associated component of the IDDM1 and its relationship to the major loci HLA-DQB1, -DQA1, and -DRB1. *Diabetes.* 2001; 50(5):1200–1205. [PubMed: 11334427]
- de Bakker PI, McVean G, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet.* 2006; 38(10):1166–1172. [PubMed: 16998491]
- Durrant C, Zondervan KT, et al. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *American Journal of Human Genetics.* 2004; 75(1):35–43. [PubMed: 15148658]
- Erlich H, Valdes AM, et al. HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families. *Diabetes.* 2008; 57(4):1084–1092. [PubMed: 18252895]
- Everitt, BS. *The Analysis of Contingency Tables.* London: Chapman and Hall; 1986.
- Forabosco P, Bouzigon E, et al. Meta-analysis of genome-wide linkage studies across autoimmune diseases. *Eur J Hum Genet.* 2009; 17(2):236–243. [PubMed: 18781189]
- Horn GT, Bugawan TL, et al. Sequence analysis of HLA class II genes from insulin-dependent diabetic individuals. *Hum Immunol.* 1988; 21(4):249–263. [PubMed: 3372263]
- Horton R, Wilming L, et al. Gene map of the extended human MHC. *Nat Rev Genet.* 2004; 5(12):889–899. [PubMed: 15573121]
- Karp DR, Marthandan N, et al. Novel sequence feature variant type analysis of the HLA genetic association in systemic sclerosis. *Hum Mol Genet.* 2010; 19(4):707–719. [PubMed: 19933168]
- Koeleman BP, Lie BA, et al. Genotype effects and epistasis in type 1 diabetes and HLA-DQ trans dimer associations with disease. *Genes Immun.* 2004; 5(5):381–388. [PubMed: 15164102]
- Li S, Khalid N, et al. Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. *Biostatistics.* 2003; 4(4):513–522. [PubMed: 14557108]

- Little AM, Parham P. Polymorphism and evolution of HLA class I and II genes and molecules. *Rev Immunogenet.* 1999; 1(1):105–123. [PubMed: 11256568]
- Liu DJ, Leal SM. Replication strategies for rare variant complex trait association studies via next-generation sequencing. *Am J Hum Genet.* 2010; 87(6):790–801. [PubMed: 21129725]
- Malkki M, Single R, et al. MHC microsatellite diversity and linkage disequilibrium among common HLA-A, HLA-B, DRB1 haplotypes: implications for unrelated donor hematopoietic transplantation and disease association studies. *Tissue Antigens.* 2005; 66(2):114–124. [PubMed: 16029431]
- Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet.* 2008; 9:387–402. [PubMed: 18576944]
- Menconi F, Osman R, et al. Shared molecular amino acid signature in the HLA-DR peptide binding pocket predisposes to both autoimmune diabetes and thyroiditis. *Proc Natl Acad Sci U S A.* 2010; 107(39):16899–16903. [PubMed: 20837527]
- Miretti MM, Walsh EC, et al. A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *Am J Hum Genet.* 2005; 76(4):634–646. [PubMed: 15747258]
- Mychaleckyj JC, Noble JA, et al. HLA genotyping in the international Type 1 Diabetes Genetics Consortium. *Clin Trials.* 2010; 7(1 Suppl):S75–87. [PubMed: 20595243]
- Nejentsev S, Howson JM, et al. Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature.* 2007; 450(7171):887–892. [PubMed: 18004301]
- Noble JA, Valdes AM. Genetics of the HLA region in the prediction of type 1 diabetes. *Curr Diab Rep.* 2011; 11(6):533–542. [PubMed: 21912932]
- Peters BA, Kermani BG, et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature.* 2012; 487(7406):190–195. [PubMed: 22785314]
- Petersdorf EW. HLA matching in allogeneic stem cell transplantation. *Curr Opin Hematol.* 2004; 11(6):386–391. [PubMed: 15548992]
- Raychaudhuri S, Sandor C, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet.* 2012; 44(3):291–296. [PubMed: 22286218]
- Rich SS, Akolkar B, et al. Results of the MHC fine mapping workshop. *Diabetes Obes Metab.* 2009; 11(Suppl 1):108–109. [PubMed: 19143823]
- Santiago JL, Li W, et al. Localization of Type 1 Diabetes susceptibility in the ancestral haplotype 18.2 by high density SNP mapping. *Genomics.* 2009; 94(4):228–232. [PubMed: 19591919]
- Schober E, Schernthaner G, et al. HLA-DR antigens in insulin-dependent diabetes. *Arch Dis Child.* 1981; 56(3):227–229. [PubMed: 7011215]
- Schork NJ, Wessel J, et al. DNA sequence-based phenotypic association analysis. *Adv Genet.* 2008; 60:195–217. [PubMed: 18358322]
- Segal MR. Regression Trees for Censored Data. *Biometrics.* 1988; 44:35–47.
- Sheehy MJ, Scharf SJ, et al. A diabetes-susceptible HLA haplotype is best defined by a combination of HLA-DR and -DQ alleles. *J Clin Invest.* 1989; 83(3):830–835. [PubMed: 2784133]
- Stewart CA, Horton R, et al. Complete MHC haplotype sequencing for common disease gene mapping. *Genome Res.* 2004; 14(6):1176–1187. [PubMed: 15140828]
- Tachmazidou I, Verzilli CJ, et al. Genetic association mapping via evolution-based clustering of haplotypes. *PLoS Genet.* 2007; 3(7):e111. [PubMed: 17616979]
- Thomson G, Marthandan N, et al. Sequence feature variant type (sfvt) analysis of the hla genetic association in juvenile idiopathic arthritis. *Pac Symp Biocomput.* 2010:359–370. [PubMed: 19908388]
- Thorsby E, Lie BA. HLA associated genetic predisposition to autoimmune diseases: Genes involved and possible mechanisms. *Transpl Immunol.* 2005; 14(3–4):175–182. [PubMed: 15982560]
- Todd JA, Bell JI, et al. HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature.* 1987; 329(6140):599–604. [PubMed: 3309680]
- Todd JA, Walker NM, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet.* 2007; 39(7):857–864. [PubMed: 17554260]

- Tzeng JY, Devlin B, et al. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *American Journal of Human Genetics*. 2003; 72(4):891–902. [PubMed: 12610778]
- Varre JS, Delahaye JP, et al. Transformation distances: a family of dissimilarity measures based on movements of segments. *Bioinformatics*. 1999; 15(3):194–202. [PubMed: 10222406]
- Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet*. 2006; 79(5):792–806. [PubMed: 17033957]
- Williams AJ, Aitken RJ, et al. Autoantibodies to islet antigen-2 are associated with HLA-DRB1*07 and DRB1*09 haplotypes as well as DRB1*04 at onset of type 1 diabetes: the possible role of HLA-DQA in autoimmunity to IA-2. *Diabetologia*. 2008; 51(8):1444–1448. [PubMed: 18504544]
- WTCCC . Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447(7145):661–678. [PubMed: 17554300]
- Yang H, Chen X, et al. Completely phased genome sequencing through chromosome sorting. *Proc Natl Acad Sci U S A*. 2011; 108(1):12–17. [PubMed: 21169219]
- Zhao LP, LeMarchand L. An analytical method for assessing patterns of familial aggregation in case-control studies. *Genetic Epidemiology*. 1992; 9(2):141–154. [PubMed: 1639245]
- Zhao LP, Li SS, et al. A method for the assessment of disease associations with single- nucleotide polymorphism haplotypes and environmental variables in case- control studies. *Am J Hum Genet*. 2003; 72(5):1231–1250. [PubMed: 12704570]

Figure 1.
The flow chart that describes the ROR Procedure

Figure 2. Recursively organizing all HLA-DRB1 alleles (outside of the ROR ring) into super variants HLA-DRB1-T1D using ROR (inside of the ring). The number 1 to 27 corresponds ROR steps. The log likelihood increments for every step are shown in the Center of the ring. The shedding colors around circle correspond to directions of associations (yellow=reference or no risk; blue=protective; pink=risk) with the density associated with the association magnitude. Line colors are indicative of the significance levels (p-values).

Figure 3. Comparison of powers between ROR (solid & red) and amino acid based association analysis (dotted & dark).

Allelic frequencies of HLA-DRB1 among cases/controls within WTCCC (discovery cohort) and T1DGC+HCT (validation cohort), and results from ROR with known associations. The rows are shaded in colors consistent to their associations as in Figure 2. HLA-DRB1*01:01 is used as the reference to be consistent with the literature.

Table 1

HLA- DRB1	Frequencies (Discovery)			ROR result		Frequencies (Validation)			Known	
	Control	%	Case	%	ROR#	CGGCTCGAC [†]	Control	%	Case	% association
*01:01	96	4.21	171	5.22	#01	-KRB----	119	4.76	492	6.54
*15:01	185	8.11	6	0.18	#02	Y----R--	351	14.05	40	0.532 Protective [1]
*14:01	67	2.94	5	0.15	#03	----CA--G	54	2.162	10	0.133 Protective [2]
*14:04	1	0.04	0	0.00		----CA--G	3	0.12	2	0.027 Protective [2]
*04:03	46	2.02	10	0.31	#04	----A-G-	18	0.721	27	0.359 Protective [2]
*04:07	22	0.96	5	0.15		----A-G-	24	0.961	8	0.106 Protective [2]
*07:01	420	18.42	158	4.82		----A-G-	336	13.45	384	5.106
*09:01	42	1.84	43	1.31		----A-G-	25	1.00	114	1.516
*01:02	18	0.79	30	0.92	#05	-----	35	1.40	43	0.572
*01:03	57	2.50	26	0.79		-----	26	1.041	62	0.824
*01:04	1	0.04	1	0.03		-----	0	0.00	1	0.013
*10:01	13	0.57	3	0.09		-----	23	0.921	9	0.12
*11:01	77	3.38	14	0.43		-----	249	9.968	44	0.585 Protective [2]
*11:02	10	0.44	5	0.15		-----	7	0.28	6	0.08 Protective [2]
*11:03	10	0.44	6	0.18		-----	11	0.44	13	0.173 Protective [2]
*11:04	22	0.96	9	0.27		-----	85	3.40	16	0.213 Protective [2]
*12:01	32	1.40	19	0.58		-----	45	1.80	30	0.40
*12:02	4	0.18	0	0.00		-----	0	0.00	0	0.00
*13:01	101	4.43	33	1.01		-----	152	6.085	87	1.157
*13:02	76	3.33	61	1.86		-----	96	3.843	196	2.606
*13:03	24	1.05	3	0.09		-----	22	0.881	10	0.133
*13:05	0	0.00	1	0.03		-----	12	0.48	0	0.00
*13:10	2	0.09	0	0.00		-----	0	0.00	1	0.013
*15:02	10	0.44	3	0.09		-----	17	0.681	5	0.066 Protective [21]
*16:01	14	0.61	10	0.31		-----	37	1.481	49	0.652
*16:02	1	0.04	0	0.00		-----	4	0.16	1	0.013

HLA- DRB1	Frequencies (Discovery)				ROR result		Frequencies (Validation)				Known association	
	Control	%	Case	%	ROR#	CGGCTCGAC [†]	Control	%	Case	%	%	%
*08:01	41	1.80	78	2.38	#06	----T---	57	2.28	189	2.51		
*08:03	3	0.13	1	0.03		----T---	6	0.24	0	0.00		
*08:04	3	0.13	3	0.09		----T---	1	0.04	4	0.05		
*08:06	0	0.00	1	0.03		----T---	0	0.00	1	0.01		
*03:01	447	19.61	1255	38.29	#07	----G---	328	13.13	2637	35.07	Risk [1]	
*03:02	0	0.00	1	0.03		----G---	0	0.00	1	0.013		
*04:01	295	12.94	894	27.27	#08	-----G-	237	9.488	2152	28.62	Risk [1]	
*04:02	8	0.35	40	1.22		-----G-	29	1.161	90	1.20	Risk [1]	
*04:04	118	5.18	292	8.91		-----G-	81	3.243	614	8.165	Risk [1]	
*04:05	14	0.61	91	2.78		-----G-	8	0.32	182	2.42	Risk [1]	
Total	2280		3278									

[†] This is the reference nucleotide sequence within HLA-DRB1 at positions 17t, 23t, 24t, 37t, 60f, 74s, 75t, 98f, 114t, where the number is the amino acid position, and f=first, s=second, and t=third is the nucleotide position in each codon. Extra nucleotide notations are for: K=G/T, R=C/T/G, Y=C/T, B=C/T/G. Note that nucleotides at position 98 and 114 are within exon 3 and hence are resulted from “imputation” described in the text.

[1] Cucca et al. 2001, [2] Koelman et al. 2004, [3] Nejentsev et al. 2007

Table 2

Disease associations with super variants with T1D for discovery and validation datasets: allelic frequencies in controls, in cases, estimated odds ratios (OR), 95% confidence intervals (CI), p-values and their corresponding sequence alleles. The rows are shaded in colors consistent to their associations as in Figure 2.

SSV	Discovery					Validation					Classical HLA-DRB1 Sequence Variants
	Control(%) (2280) [†]	Case(%) (3278)	OR	95% CI	p-value	Control(%) (2498)	Case(%) (7520)	OR	95% CI	p-value	
#01	4.21	5.22	1.00	---	---	4.76	6.54	1.00	---	---	01:01
#02	8.11	0.18	0.02	(0.01, 0.06)	< 2E-16	14.05	0.53	0.03	(0.02, 0.04)	< 2E-16	15:01
#03	2.98	0.15	0.05	(0.02, 0.14)	2.83E-09	2.28	0.16	0.07	(0.05, 0.10)	3.21E-13	14:01, 14:04
#04	23.25	6.59	0.25	(0.18, 0.34)	< 2E-16	16.13	7.09	0.35	(0.3, 0.41)	2.17E-13	04:03, 04:07, 07:01, 09:01
#05	20.70	6.83	0.26	(0.19, 0.36)	4.75E-16	32.87	7.62	0.19	(0.16, 0.21)	< 2E-16	01:02, 01:03, 01:04, 10:01*
#06	2.06	2.53	1.03	(0.63, 1.67)	0.92	2.56	2.58	0.94	(0.76, 1.16)	0.77	08:01, 08:03, 08:04, 08:06
#07	19.61	38.32	1.56	(1.16, 2.09)	3.24E-03	13.13	35.08	2.11	(1.84, 2.42)	3.88E-08	03:01, 03:02
#08	19.08	40.18	2.08	(1.54, 2.82)	2.19E-06	14.21	40.40	2.73	(2.38, 3.14)	4.09E-13	04:01, 04:02, 04:04, 04:05

[†]Number of haplotypes

* Other alleles are HLA-DRB1* 11:01, 11:02, 11:03, 11:04, 12:01, 12:02, 13:01, 13:02, 13:03, 13:05, 13:10, 15:02, 16:01, 16:02

Table 3

Comparisons of results from Raychaudhuri's approach and from ROR obtained on T1D and HLA-DRB1 (WTCCC dataset). A total of four amino acids (position 67, 74, 86 and 96) are chosen, and form 24 different alleles, associating with a class of HLA-DRB1 alleles (far right). Their odds ratios, confidence intervals and p-values, together with their allelic frequencies are shown in column 5–9. Corresponding super variants (ROR#01 to #08) are shown in the tenth column.

Amino Acids (position)		Raychaudhuri's Approach			Allele Freq		ROR			Classical HLA-DRB1 alleles
67	74	86	96	OR	95% CI	Pr(> z)	Controls	Cases	ROR	
L	A	G	E	1.00	---	---	4.21	5.22	#01	*01:01
I	A	V	Q	0.02	(0.01, 0.06)	<2e-16	8.11	0.18	#02	*15:01
L	E	V	H	0.05	(0.02, 0.14)	3.62E-09	2.98	0.15	#03	*14:01, *14:04
L	E	V	Y	0.16	(0.07, 0.34)	2.64E-06	2.02	0.31	#04	*04:03
L	E	G	Y	0.19	(0.07, 0.56)	0.00256	0.96	0.15		*04:07
I	Q	G	H	0.22	(0.15, 0.30)	<2e-16	18.42	4.82		*07:01
F	E	G	H	0.70	(0.40, 1.22)	0.21066	1.84	1.31		*09:01
F	A	G	H	0.11	(0.06, 0.20)	1.47E-11	3.38	0.46	#05	*11:01, *13:05
L	A	G	Q	0.12	(0.03, 0.45)	0.00178	0.61	0.09		*10:01, *16:02
I	A	V	H	0.21	(0.13, 0.32)	6.08E-13	6.36	1.74		*11:02, *12:01, *13:01, *13:10
F	A	V	H	0.23	(0.12, 0.48)	6.01E-05	1.58	0.45		*11:03, *11:04, *12:02
I	A	G	E	0.24	(0.13, 0.42)	1.45E-06	2.50	0.79		*01:03
I	A	G	Q	0.24	(0.06, 1.07)	0.06181	0.44	0.09		*15:02
I	A	G	H	0.35	(0.23, 0.56)	6.31E-06	4.38	1.95		*13:02, *13:03
F	A	G	Q	0.55	(0.20, 1.46)	0.23003	0.61	0.31		*16:01
L	A	V	E	0.72	(0.36, 1.45)	0.3571	0.83	0.95		*01:02, *01:04
I	L	G	H	0.24	(0.02, 2.84)	0.25874	0.13	0.03	#06	*08:03
F	L	G	H	1.09	(0.65, 1.83)	0.73485	1.80	2.38		*08:01
F	L	V	H	1.37	(0.22, 8.62)	0.73754	0.13	0.12		*08:04, *08:06
L	R	V	H	1.61	(1.19, 2.18)	0.00214	19.61	38.29	#07	*03:01
L	R	G	H	>1.00	NA	NA	0.00	0.03		*03:02
L	A	V	Y	1.37	(0.96, 1.97)	0.08314	5.18	8.91	#08	*04:04
L	A	G	Y	2.43	(1.75, 3.38)	1.03E-07	13.55	30.05		*04:01, *04:05
I	A	V	Y	4.54	(1.83, 11.24)	0.00108	0.35	1.22		*04:02