Hindawi Publishing Corporation
International Journal of Genomics
Volume 2015, Article ID 256818, 13 pages
http://dx.doi.org/10.1155/2015/256818

Research Article

Genetic Vectors as a Tool in Association Studies: Definitions and Application for Study of Rheumatoid Arthritis

Igor Sandalov[1,2] and Leonid Padyukov[1]

[1] *Rheumatology Unit, Department of Medicine Solna, Karolinska Institutet, CMM L8:04, 17176 Stockholm, Sweden*
[2] *L.V. Kirensky Institute of Physics, Akademgorodok 50, Krasnoyarsk 660036, Russia*

Correspondence should be addressed to Leonid Padyukov; leonid.padyukov@ki.se

Received 13 November 2014; Accepted 6 February 2015

Academic Editor: Ian Dunham

To identify putative relations between different genetic factors in the human genome in the development of common complex disease, we mapped the genetic data to an ensemble of spin chains and analysed the data as a quantum system. Each SNP is considered as a spin with three states corresponding to possible genotypes. The combined genotype represents a multispin state, described by the product of individual-spin states. Each person is characterized by a single genetic vector (GV) and individuals with identical GVs comprise the GV group. This consolidation of genotypes into GVs provides integration of multiple genetic variants for a single statistical test and excludes ambiguity of biological interpretation known for allele and haplotype associations. We analyzed two independent cohorts, with 2633 rheumatoid arthritis cases and 2108 healthy controls, and data for 6 SNPs from the HTR2A locus plus shared epitope allele. We found that GVs based on selected markers are highly informative and overlap for 98.3% of the healthy population between two cohorts. Interestingly, some of the GV groups contain either only controls or only cases, thus demonstrating extreme susceptibility or protection features. By using this new approach we confirmed previously detected univariate associations and demonstrated the most efficient selection of SNPs for combined analyses for functional studies.

1. Introduction

The amount of data from genetic studies is growing and its structure is becoming more complex due to the rapid development of new genotyping and sequencing techniques. However, current understanding of the correlation between found genotypes and observed phenotypes remains an obstacle for more efficient use of these data in medicine and biology. There are only a few methods for integration of genetic data in a reasonable way for statistical and functional analyses. An important feature of genetic data is that it is not a collection of random variations, but rather a system with intrinsic correlation between genetic variants. Investigations of study populations are often based on the frequencies of alleles rather than genotypes. However, these variants are distributed on two parallel carriers of information, paired sets of human chromosomes, and in most cases both alleles contribute to the phenotype. Traditionally, statistical analysis in genetics is based on several simplified models with limited options for integration of multiple variants in the same analysis. Although in recent years an exponential growth of different approaches to search for epistatic effects, or gene-gene interaction, in studies of complex traits is observed (see review [1]), a simultaneous consideration of multiple variants remains difficult due to "bottom-up" the design of the strategy of the search.

In this study we suggest and evaluate a genetic vector approach (GVA; briefly described in [2]) that would allow integration of available genotyping data in an unambiguous way, selection of the most representative combinations for the statistical analyses, and, finally, significant reduction of the number of variants that need to be inspected for functional studies. Our approach provides possibility to combine the advantage of joining several genetic markers in one combination (haplotype) with consideration to genotype (not allele) for each marker. Instead of ambiguous connection of allele or haplotype to certain phenotype in each individual, in GVA each individual has only a single genetic vector that corresponds to the phenotype. Thus the genotype-phenotype

relation is unambiguous and it does not need *a priori* biological interpretation (dominance/recessiveness). GVA also permits expression of the traditional measures of the association for separate genotypes and, particularly, enables detection of gene-gene interaction.

2. Methods

2.1. Genetic Vectors for Association Studies. Due to the complexity of genetic data, multiple options for the analyses are available. However, the majority of genetic analyses are based on univariate tests for independent genetic markers and applications with simultaneous employment of several genetic markers are relatively rare. Based on linkage disequilibrium, haplotypes are considered to be a valuable tool [3], but this approach does not provide unambiguous assessment to phenotype, since individuals with heterozygotic state could be, at best, considered as an intermediate group or should be combined with homozygotic state in the absence of a biological hypothesis. We constructed an approach for genetic analyses that integrates a number of markers as a whole entity, regardless of the specific linkage on a particular chromosome.

The design of the model is based on integration of different genetic variants in a reasonable biological group by assigning each individual to a sequence of genotypes in relation to given genetic variants. Since the combinations available in the population are not random and reflect the population's genetic structure, the number of these sequences of genotypes is significantly less than the number of individuals in a study population of reasonable size. The individuals identical in the order of genotypes represent a group, which may have differential representation of phenotypes. In a case-control design the frequencies of these groups could be statistically compared to identify those that are most likely to associate with a selected phenotype.

Imagine that we are interested in the degree of association to a certain phenotype of several single-nucleotide polymorphisms (SNPs), which may belong to the same, or to different chromosomes. A polymorphism is defined at the level of the whole study population, whereas for a particular person each SNP has a definite genotype value. Let us make the following convention: we choose and fix the order of the SNPs. Then each person will be characterized by the set of genotypes that can be thought of as a *genetic vector* (GV), characterizing a particular person. The number of genetic variations included in this vector is the length of the GV. The study population is then characterized by the set of GVs.

In an association study, each GV ν_i "contains" some number $n_{h,i}$ of controls and $n_{s,i}$ cases. Then the structure of the study population is described by the Table 1.

Thus, this is a simple way to avoid working with separate alleles, and even with genotypes; instead, the focus lies on their relevant combinations in the study population, that is, on the genetic vectors. One of the advantages worth mentioning is that if there exists some, say, genotype-genotype interactions, these will be taken into account automatically, since at this first stage the approach does not require any assumptions about the statistical nature of the GV's constituents: this is simply one of the possible ways to represent the experimental data set. Since we now know the number of healthy controls and number of cases, as well as the total numbers of both, we can apply standard statistical machinery, testing the statistical hypotheses for whole GVs; particularly, in order to discover the "most promising" GVs, we can (and will) evaluate the odds ratios, the regions with the greatest fluctuations on the basis of, say, Fisher's exact tests, and so on.

Table 1: Structure of the study population in terms of GVs: each GV $|\nu_i\rangle$ is "populated" by n_h controls and n_s cases.

Type of GV	n_h	n_s	GV
ν_1	$n_{h,1}$	$n_{s,1}$	$\|\nu_1\rangle$
ν_2	$n_{h,2}$	$n_{s,2}$	$\|\nu_2\rangle$
⋮	⋮	⋮	⋮
ν_N	$n_{h,N_{\mathrm{GV}}}$	$n_{s,N_{\mathrm{GV}}}$	$\|\nu_{N_{\mathrm{GV}}}\rangle$

The GVA consists of several stages:

(1) selection of genetic markers,

(2) genotyping for two study populations,

(3) assigning of genetic vector value to each individual,

(4) sorting all individuals in the study population into groups, according to particular GVs and chosen phenotype parameters,

(5) exclusion of GVs with zero number of individuals in any subgroup (subset),

(6) computation of odds ratio and relative risk (OR/RR), pathogenic, and protective genetic vectors,

(7) sorting GVs according to statistical confidence,

(8) exclusion of GVs with low confidence,

(9) analysis of the match between study populations: focus on replication and absence of opposite effect for the vector.

The mathematical details of the GVA itself are given below (Section 2.3–2.5).

2.2. Experimental Data Sets and GV Definitions. To demonstrate the capacity of the methodology we applied it to the analyses of data from two study populations: from the Swedish Epidemiological Investigation of Rheumatoid Arthritis (EIRA, 1820 cases and 947 controls) and from the North American Rheumatoid Arthritis Consortium (NARAC, 813 ACPA-positive cases and 1161 controls).

The EIRA study population is a population-based case-control study of incident cases of rheumatoid arthritis (RA) in which all patients fulfilled the American College of Rheumatology (ACR) 1987 criteria [4]. Controls were randomly selected from the Swedish national population registry, taking into consideration the patient's age, sex, and residential area. More details about the EIRA study population have been described elsewhere [5].

The cases in the NARAC study population consisted of RA patients of self-reported white ancestry, who were randomly drawn from four different groups of patients, while

Table 2: Encoding of SNP states: 1 = rs6314; 2 = rs977003; 3 = rs1328674; 4 = rs2070037; 5 = rs6313; 6 = rs6311. Each SNP can aquire any of genotype values listed in its column. Each GV is a product of SNP states in the order given in the table. Columns correspond to reference sequence number of SNP or to shared epitope (SE) genotype.

SNP	$\|1,\gamma_1\rangle$	$\|2,\gamma_2\rangle$	$\|3,\gamma_3\rangle$	$\|4,\gamma_4\rangle$	$\|5,\gamma_5\rangle$	$\|6,\gamma_6\rangle$	$\|\mathrm{SE},\gamma_7\rangle$
$\gamma_i = 1$	CC	AA	CC	CC	AA	CC	No
$\gamma_i = 2$	CT	AC	CT	CT	AG	CT	Single
$\gamma_i = 3$	TT	CC	TT	TT	GG	TT	Double

controls were recruited from the New York Cancer Project [6, 7]. Both studies were conducted after obtaining approval from the Regional Ethics Committees and in accordance with the Declaration of Helsinki.

Below, we apply GVA for the analysis of EIRA's and NARAC's association to rheumatoid arthritis (RA) using the data for 6 SNPs from the HTR2A locus, rs6314, rs977003, rs1328674, rs2070037, rs6313, and rs6311 and for HLA-DRB1 shared epitope alleles. We defined SE alleles as any of HLA-DRB1*01 (not *0103), *04 and *10; previous study suggested very low frequency for non-SE HLA-DRB1*04 variants in Caucasian populations [8]. Conventional evaluation for HTR2A genetic markers was performed previously [9].

We defined the GV states as a product of SNP states, as shown in Table 2. In order to compare the GV structure between the EIRA and NARAC study populations and to perform statistical evaluation, we, first, assigned individuals to a particular GV group using individual genotypes and, second, normalized the number of people, for each GV (according to (10) below).

2.3. Mathematical Details of the Genetic Vector Approach (GVA). Let us consider a population, consisting of N_h controls (h = healthy) and N_s (s = sick) cases. The number of genetic markers is a matter of preliminary research and will be assigned as N_{SNP}. We choose here SNPs for simplicity, but the approach can be easily extended for multiallelic markers. There are no restrictions on the position of markers within chromosomes, and variants from different chromosomes could be used. In the whole study population, with the dizygotic human genome, each SNP can normally produce 3 different categorical values, or genotypes. However, these categories are exclusive and only one of the categories can be found in the DNA of an individual person. This is similar to one of the components (say, z) of spin $S = 1$, which can acquire three values, $S^z = 1, 0, -1$, and could be described by corresponding three states, $|1\rangle, |0\rangle, |-1\rangle$. The description of the study population with a set of SNPs is similar to the description in physics of an ensemble of spin chains (see, e.g, [10]). Each of the spin chains contains N_{SNP} spins and can be in different quantum states. The latter is described by the set of vectors:

$$|\nu_i\rangle = |S_{1i}\rangle\,|S_{2i}\rangle\,|S_{3i}\rangle\cdots\left|S_{N_{\mathrm{SNP}},i}\right\rangle|\Upsilon\rangle\,, \tag{1}$$

where Υ represents other variables; the spin corresponds to the SNP, and its values correspond to the genotypes. Therefore, a combination of genotypes form a vector for each person with known genotypes and we refer to these combinations as *genetic vectors* (GVs).

Different statistical characteristics of the study population, such as the average frequency of people with a certain genotype and correlations between genotypes and phenotypes, can be obtained by, first, averaging certain spin-operator combinations on these multispin states and, second, averaging the result over the population of interest. However, it is more convenient for us to speculate not in terms of $\widehat{S}^z$-operators, but, rather, in terms of the operators, defined as follows:

$$\begin{aligned}\widehat{X}^{\gamma}_{\nu i} &= |\nu,i,\gamma\rangle\,\langle\nu,i,\gamma|\,;\\ \widehat{X}^{\gamma}_{\nu i}\,|\nu_1,i_1,\gamma_1\rangle &= \delta_{\nu\nu_1}\delta_{ii_1}\delta_{\gamma\gamma_1}\,|\nu,i,\gamma\rangle\,.\end{aligned} \tag{2}$$

These variables can be interpreted as the operators of population numbers (PNs) of the γ-state in the cell i of the genetic vector ν. The X-operators are projection operators: $(X^{\gamma}_{\nu i})^2 = X^{\gamma}_{\nu i}$, $\gamma = 1, 2, 3$. Notice that the projection operators are sometimes used for description of spin systems, since the spin operators can be easily expressed in terms of $X^{\gamma}_{\nu i}$-operators. For example, z-projecture of spin operator is $\widehat{S}^z_{\nu i} = \sum_{\gamma} m_{\gamma} X^{\gamma}_{\nu i} = 1\cdot X^1_{\nu i} + 0\cdot X^2_{\nu i} + (-1)\cdot X^3_{\nu i} = X^1_{\nu i} - X^3_{\nu i}$; that is, it is just the difference between the population-number operators of two states.

In genetics this multiple-spin state, describing the combination of genotypes, GV, and characterizing each person, could be obtained from the experimental genotyping. Then, one can compare all expected values of interest in groups via the expectation values of corresponding operators on the wave functions (GVs) for two ensembles, for example, controls and cases. We are dealing here with a static problem: all the states of "spin chains" are given; they do not depend on time; the total wave function $|\Psi^{\alpha}\rangle$ of each of the ensembles α = ctrl, case is a product of the wave functions of separate "chains," $\prod_{\nu}|\nu\rangle$. The expectation value of the population-number operator for a certain person is

$$\begin{aligned}N^{\alpha}_{\nu i\gamma} &\equiv \left\langle\Psi^{\alpha}\right|X^{\gamma}_{\nu i}\left|\Psi^{\alpha}\right\rangle = \langle\nu|\,X^{\gamma}_{\nu i}\,|\nu\rangle\\ &= \begin{cases}1 & \text{if } (i,\gamma)\in \mathrm{GV}\,|\nu\rangle\,;\\ 0 & \text{otherwise.}\end{cases}\end{aligned} \tag{3}$$

Therefore, in general, the individual ν is characterized by $3N_{\mathrm{SNP}}$ numbers ($i = 1,\ldots,N_{\mathrm{SNP}}$, and $\gamma = 1, 2, 3$, if other parameters are excluded) or by the matrix $3\times N_{\mathrm{SNP}}$ of PNs for each ν. We will refer to it as *index matrix* (IM):

$$\widehat{C}_{\nu} \equiv \left\|N_{\nu i\gamma}\right\|. \tag{4}$$

Thus, an individual is fully characterized by this matrix as described above. Each column, describing the filling of the "spin-one state", can contain only one nonzero value equal to one and two others equal to zero. Each column represents a particular variation and, for a SNP, has three levels defined by nucleotides at the SNP position. The matrix for selected genotypes is a matter of experimental design and may reflect either a set of selected SNPs of interest or a sequence of all SNPs at a certain locus. For example, if we are interested in an ensemble of three SNP chains, with $\mathrm{SNP}_1 = A$, $\mathrm{SNP}_2 = B$, and $\mathrm{SNP}_3 = C$, with $A = a_1, a_2, a_3$, $B = b_1, b_2, b_3$, and $C = c_1, c_2, c_3$, where a_i, b_j, c_k are corresponding genotypes, then some people may be in the state characterized by the genetic vector:

$$\left|\nu_0\right\rangle = \left|a_2\right\rangle_1 \left|b_1\right\rangle_2 \left|c_3\right\rangle_3 . \tag{5}$$

Corresponding to this GV index matrix is

$$C_{\nu_0} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{6}$$

These IMs will be used in our further calculations. For statistical evaluation of differences between controls and cases we will develop two matrices $3 \times N_{\mathrm{SNP}}$ of PNs:

$$N_{i\gamma}^{\alpha} = \frac{1}{N_\alpha} \sum_{\nu} \left\langle \Psi^{\alpha} \right| X_{\nu i}^{\gamma} \left| \Psi^{\alpha} \right\rangle , \qquad \alpha = h \ \ (\text{i.e., ctrl}) , \ \ s \ \ (\text{i.e., case}) . \tag{7}$$

These, in fact, are the genotype-frequency maps for controls and cases for the study population of interest.

In the same fashion we can obtain the rate for people, who have a pair of certain genotypes (i, γ) and (i', γ'); this rate can be calculated from the expression:

$$P_{i,\gamma;i',\gamma'}^{\alpha} = \frac{1}{N_\alpha} \sum_{\nu} \left\langle \Psi^{\alpha} \right| X_{\nu i}^{\gamma} X_{\nu i'}^{\gamma'} \left| \Psi^{\alpha} \right\rangle , \quad \alpha = h, s. \tag{8}$$

The expression for triplets contains a product of three X-operators and so on.

Each GV ν_i "contains" some number $n_{h,i}$ of controls and $n_{s,i}$ cases, whereas the structure of the study population is described by Table 1.

Notice that the set of GVs is orthonormalized

$$\left\langle \nu_i \mid \nu_j \right\rangle = \delta_{ij} = \begin{cases} 1, & i = j; \\ 0, & i \neq j \end{cases} \tag{9}$$

and, therefore, can be considered as a basis set in N_{GV}-dimensional space, whereas each part of the study population, controls and cases, can be presented as a point in this space. If some other study population contains GVs that are not included in the set $|\nu_1\rangle, |\nu_2\rangle, \ldots, |\nu_{N_{\mathrm{GV}}}\rangle$, the set should be complemented by these GVs in order to obtain a complete set for the extended population, including all populations in question.

2.4. Statistical Treatment of GVs. In order to avoid possible confusion it is important to remember that each GV describes the set of SNP values. If the population has a group of individuals with the same GV, we can use this fact in the statistical treatment of the population, describing the whole population in terms of these larger-scale variables, GVs. Then, each GV within a given study population is characterized by the number of healthy controls N_h and the number of cases N_s and we can introduce two frequencies:

$$p_\nu^\alpha = \frac{n_\nu^\alpha}{N_\alpha}, \quad \alpha = h, s. \tag{10}$$

Comparing these, we can determine which of the GVs are more protective and which are more pathogenic. As a criterion for the separation of GV subgroups with protective and pathogenic behavior we choose here the odds ratio, defined by GV frequencies:

$$\mathrm{OR}_\nu^{(\mathrm{GV})} = \frac{p_\nu^s / \left(1 - p_\nu^s\right)}{p_\nu^h / \left(1 - p_\nu^h\right)}. \tag{11}$$

The difference between the standard definitions of OR and the one used here is discussed in the book [10, page 154]. Further, we have to choose some threshold $\mathrm{OR}_{\mathrm{tr}}$, which reflects a chosen criterion of statistical confidence, and select all GVs that satisfy the condition

$$\mathrm{OR}_\nu^{(\mathrm{GV})} > \mathrm{OR}_{\mathrm{tr},\nu}^{s} \tag{12}$$

for "pathogenic" GVs and for the "protective" ones, for which $\mathrm{OR}_\nu^{(\mathrm{GV})} < \mathrm{OR}_{\mathrm{tr},\nu}^{h}$. The standard obvious threshold in (12) would be $\mathrm{OR}_{\mathrm{tr},\nu} = \mathrm{OR}_{\mathrm{tr},\nu}^{h} = \mathrm{OR}_{\mathrm{tr},\nu}^{s} = 1$. However, statistical fluctuations in the vicinity of the threshold $\mathrm{OR}_{\mathrm{tr},\nu} = 1$ may make the choice of the groups, described by certain GVs, doubtful. In order to garantee that the selected GVs remain on a safe side, one can choose a harder criteria of the GV selection. Of course, the threshold for "protective" GVs $\mathrm{OR}_{\mathrm{tr},\nu}^{h}$ not obligatory should be chosen equal to $1/\mathrm{OR}_{\mathrm{tr},\nu}^{s}$. Due to the different number of people in each GV, the threshold value $\mathrm{OR}_{\mathrm{tr},\nu}$ can occur as ν-dependent. Let us denote the subset of GVs defined by the inequality (12) as Ω_{trust} (the details of this subset will be discussed in Sections 2.5 and 3.2). Then the combination of variants, responsible for a large OR, can be visualized by summing the IMs of the GVs that fulfil (12). This summation produces two matrices,

$$\overline{C_{\Omega_{\mathrm{crit}}}^{h}} = \frac{1}{N_h} \sum_{\nu} C_{\nu,\Omega_{\mathrm{trust}}}^{h}, \qquad \overline{C_{\Omega_{\mathrm{crit}}}^{s}} = \frac{1}{N_s} \sum_{\nu} C_{\nu,\Omega_{\mathrm{trust}}}^{s}, \tag{13}$$

which represent frequencies of genotypes in selected control and case groups. The cells in the matrix $\Delta\overline{C_{\Omega_{\mathrm{trust}}}} = \overline{C_{\Omega_{\mathrm{trust}}}^{s}} - \overline{C_{\Omega_{\mathrm{trust}}}^{h}}$ that contain the largest difference give us a possible indication of which of the genotypes, entering different GVs, are responsible for the large ORs displayed by the GVs.

2.5. The Choice of Thresholds. Let us now discuss a criterion for the choice of thresholds $\mathrm{OR}_{\mathrm{tr},\nu}$. In the first step, all N_t individuals in the population are sorted in terms of corresponding

Table 3: Contingency table for a GV.

	Belongs to GV ν	Does not belong to GV ν
s	n_ν^s	$N_s - n_\nu^s$
h	n_ν^h	$N_h - n_\nu^h$

genetic vectors (GV) of a length N_{GV}. Each GV ν contains n_ν^h of healthy controls (HCs) and n_ν^s cases, with a total of $n_\nu^t = n_\nu^{\mathrm{ctrl}} + n_\nu^{\mathrm{case}}$ individuals in each GV. The whole population consists of N_h controls and N_s case individuals, with the total number of individuals in the population $N_t = N_h + N_s$. Then for each GV ν we can write down the contingency Table 3 and the odds ratio

$$\mathrm{OR}_\nu\left(n_\nu^s\right) = \frac{n_\nu^s\left(N_h - n_\nu^h\right)}{n_\nu^h\left(N_s - n_\nu^s\right)}, \tag{14}$$

based on this table. The full set of GVs can be sorted into two subsets, one with OR > 1, which reflects the pathogenic combinations, and the rest with OR < 1, corresponding to the protective combinations of genotypes.

At this stage we have to estimate the degree of confidence for each OR_ν. In order to do this, we introduce the random variable $n_{\nu k}$, which describes a number of case individuals, selected randomly from n_ν^t persons, with the number n_ν^t *kept fixed*. Then the probability to pick up precisely this number for fixed values n_ν^t, N_h, and N_s is described by the hypergeometric distribution (as used in Fisher's exact test):

$$\frac{\binom{N_s}{n_{\nu k}^s}\binom{N_h}{n_{\nu k}^h}}{\binom{N_t}{n_\nu^t}}, \quad n_{\nu k} = 0, 1, 2, \ldots, n_\nu^t, \tag{15}$$

where $\binom{N}{n}$ are binomial coefficients. Constructing the complex random variable of interest, OR, for the chosen $n_{\nu k}$,

$$\mathrm{OR}_{\nu k}^{\mathrm{rand}} = \frac{n_{\nu k}\left(N_h - n_\nu^t + n_{\nu k}\right)}{\left(n_\nu^t - n_{\nu k}\right)\left(N_s - n_{\nu k}\right)}, \tag{16}$$

we can calculate for each GV the average odds ratio

$$\mu_1(\mathrm{OR}) = M\left[\mathrm{OR}_{\nu k}^{\mathrm{rand}}\right] = \sum_{k=1}^{n_\nu^t - 1} \frac{n_{\nu k}\left(N_h - n_\nu^t + n_{\nu k}\right)}{\left(n_\nu^t - n_{\nu k}\right)\left(N_s - n_{\nu k}\right)} p\left(n_{\nu k}\right), \tag{17}$$

and, say, variance, $\mu_2(\mathrm{OR}) = \mathrm{var}[\mathrm{OR}_{\nu k}^{\mathrm{rand}}] = M[(\mathrm{OR}_{\nu k}^{\mathrm{rand}})^2] - (M[\mathrm{OR}_{\nu k}^{\mathrm{rand}}])^2$. This would be sufficient if we had normal distribution of the variable of interest, $\mathrm{OR}_{\nu k}^{\mathrm{rand}}$, but our case is far from this. One could also try to use the higher moments in order to build the confidence intervals for $\mathrm{OR}_\nu^{\mathrm{rand}}$ for each GV ν. However, due to asymmetry of the distribution for the odds ratio $\mathrm{OR}_{\nu k}^{\mathrm{rand}}$ and the presence of GVs with a small total number n_ν^t of individuals in GV, this method is impractical. Instead, we build the distributions for the complex random variable $\mathrm{OR}_{\nu k}^{\mathrm{rand}}$, $p(\mathrm{OR}_{\nu k}^{\mathrm{rand}})$, and find the thresholds $(\mathrm{OR}_{\mathrm{lower}}^{\mathrm{tr}}, \mathrm{OR}_{\mathrm{upper}}^{\mathrm{tr}})$ for the interval, where $\mathrm{OR}_{\nu k}^{\mathrm{rand}}$ experiences the largest fluctuations (for the normal distribution this would be the interval $[-\eta\sigma, \eta\sigma]$, where σ is the standard deviation and $[1 - \mathrm{erf}(\eta/\sqrt{2})]/2$ is chosen accuracy) directly from the equation based on this distribution:

$$\begin{aligned} \alpha_{\nu,\mathrm{upper}} &= \sum_{\mathrm{OR}_{\nu k}^{\mathrm{rand}} > \mathrm{OR}_{\mathrm{upper}}^{\mathrm{tr}}} p\left(\mathrm{OR}_{\nu k}^{\mathrm{rand}}\right), \\ \alpha_{\nu,\mathrm{lower}} &= \sum_{\mathrm{OR}_{\nu k}^{\mathrm{rand}} < \mathrm{OR}_{\mathrm{lower}}^{\mathrm{tr}}} p\left(\mathrm{OR}_{\nu k}^{\mathrm{rand}}\right). \end{aligned} \tag{18}$$

We choose a 5% level of accuracy, that is, $\alpha_{\nu,\mathrm{upper/lower}} = 0.05$. Since the full number of steps in the discrete distribution $p(n_{\nu k}; n_\nu^t, N_s, N_h)$ for the total number of individuals in GV n_ν^t is restricted by n_ν^t, the best accuracy α that can be achieved is also restricted by the magnitude of this step; the sum in (18) will consist of only one term with this minimal possible step. As normal, a decrease of α also decreases the number of contributing GVs, decreasing the total statistical power.

The other way to estimate the degree of confidence for the experimentally found $\mathrm{OR}_{\nu,\mathrm{exp}}$ for each GV is to evaluate the sum of probabilities for all possible $\mathrm{OR}_{\nu k}^{\mathrm{rand}}$ exceeding $\mathrm{OR}_{\nu,\mathrm{exp}}$:

$$P_{\nu,\mathrm{CI}} = \sum_{\mathrm{OR}_{\nu k}^{\mathrm{rand}} > \mathrm{OR}_{\nu,\mathrm{exp}}} p\left(\mathrm{OR}_{\nu k}^{\mathrm{rand}}\right). \tag{19}$$

The criteria are illustrated in Figure 1.

We use both criteria.

The GVs, which contain either only cases ($n_\nu^h = 0$, $n_\nu^s \neq 0$, *i.e.*, "solely sick") or only controls ($n_\nu^h \neq 0$, $n_\nu^s = 0$, *i.e.*, "solely healthy"), require separate consideration. For simplicity we call them "zero GVs."

3. Results and Discussion

3.1. GV Frequency Distributions. We have found that all individuals in two study populations, based on selected SNPs from HTR2A and HLA-DRB1 SE alleles for the chosen GV length (number of markers), can be ascribed to 161 GVs for EIRA and 163 GVs for NARAC; 131 of them are common GVs for both studies. The total basis of GVs for these both cohorts, therefore, consists of 193 GVs, while Figure 2 displays the results for common GVs only.

The rest of the GVs contain a relatively small number of individuals in each of the cohorts: 40 people of 2767 total in 30 GVs (1.45%) in EIRA and 36 people of 1974 total in 32 GVs (1.82%) in NARAC. Since each of these noncommon GVs contains just one or two people, it is obvious that these relatively rare GVs would be unlikely to give a significant contribution to the statistical characteristics of the cohorts. Thus, the vast majority of GVs in both cohorts are the same and it is reasonable to compare GV frequencies between the two study populations.

The data in Figure 2 demonstrate the high level of coherency of the GV distribution within each study population (EIRA and NARAC) and provides a rationale for statistical comparison between RA cases and controls. It is also evident that these two populations differ in the frequencies

Table 4: The number of GVs in different groups: $N^{\text{all}}_{\text{GV}}$ is total number of GVs, h means "healthy controls," s stands for cases ("sick"), and the index "with Zeros" denotes the GVs that do not contain individuals in one of the groups, h or s. These types of GVs are called zero GVs; "no Zeros" indicate that these sets of GVs do not contain zero GVs. Obviously, $N^{h}_{\text{GV,noZeros}} + N^{h}_{\text{GV,withZeros}} = N^{s}_{\text{GV,noZeros}} + N^{s}_{\text{GV,withZeros}} = N^{\text{all}}_{\text{GV}}$.

	$N^{\text{all}}_{\text{GV}}$	$N^{h}_{\text{GV,noZeros}}$	$N^{s}_{\text{GV,noZeros}}$	$N^{h}_{\text{GV,withZeros}}$	$N^{s}_{\text{GV,withZeros}}$	$N^{\text{Common}}_{\text{GV,noZeros}}$
GV_{EIRA}	161	109	146	52	15	62
GV_{NARAC}	163	135	101	28	62	62

Table 5: Numbers of vectors and individuals after separation of GVs specific only for one population and with zero frequency in any of subsets: $N^{\text{Common}}_{\text{GV,noZeros}}$ and $N^{\text{allCommon}}_{\text{GV}}$ are the numbers of common GVs in both sets, $n^{\alpha}_{\text{noZeros}}$ is number of people after, and n^{α}_{tot} is number of individuals before cleaning, index $\alpha = h, s$ ($E \equiv$ EIRA, $N \equiv$ NARAC, h means "healthy controls", s stands for cases ("sick")).

	$N^{\text{Common}}_{\text{GV,noZeros}}$	$N^{\text{allCommon}}_{\text{GV}}$	n^{h}_{noZeros}	n^{h}_{tot}	n^{s}_{noZeros}	n^{s}_{tot}	n^{h+s}_{noZeros}	n^{h+s}_{tot}
E	62	131	779 (83%)	938	1500 (83%)	1789	2279 (83%)	2727
N	62	131	866 (76%)	1134	711 (88%)	804	1577 (81%)	1938

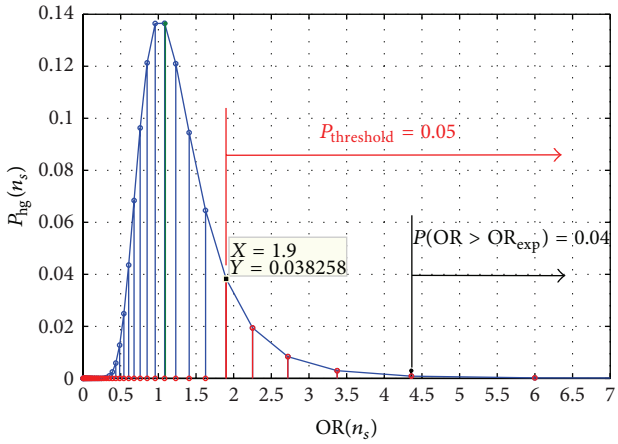

Figure 1: The example of the OR-distribution at fixed n^{t}_{ν} : the density of probability for hypergeometric distribution $P_{hg}(n_s)$ versus odds ratio $\text{OR}(n_s)$. As seen, the distribution is asymmetric. The plot is built for the following parameters: $N_s = 1500$; $N_h = 779$; $n^{t}_{\nu} = 37$. The expectation value $\mu_1(\text{OR}^{\text{rand}}_{\nu k}) = 1.0929$ (shown by green line), the threshold OR, $\text{OR}^{\text{tr}}_{\text{upper}} = 1.9$. This is the solution to (18), meaning that the sum of probabilities from $P(\text{OR}^{\text{tr}}_{\text{upper}})$ till $P(\text{OR}_{\text{max}})$ (shown by red vertical lines) is equal to $\alpha = 0.05$. At the end, the probability of finding OR greater than the experimental value of the odds ratio, $\text{OR}_{\nu,\text{exp}} = 4.3584$, equals 0.04; the value $\text{OR}_{\nu,\text{exp}}$ is taken from experiments, described above, and is shown by a black star ($*$).

of specific GVs, which is likely to reflect a different genetic backgrounds or differences in the selection criteria. More specifically, as can be seen from Figure 2(c), frequencies of several GVs for the RA cases in the two studies differ significantly, with some GVs absent from NARAC, while relatively common in EIRA.

3.2. Stratification of the Common GVs and an Analysis of the Consistency. In order to facilitate identification and comparison between high-risk and protective GVs in the two study populations, we selected GVs which display $\text{OR} > \text{OR}^{\text{thr}}_{\text{upper}}$ or $\text{OR} < \text{OR}^{\text{thr}}_{\text{lower}}$ (see (18) in Section 2.5). However, the GVs with either $n^{\text{ctrl}}_{\nu} = 0$ or $n^{\text{case}}_{\nu} = 0$ should be considered separately, since in this case OR cannot be formally defined. These subgroups of GVs with zero values will be described in Section 3.4. The distribution of GVs between the groups is shown in Table 4.

After exclusion of the GVs specific only for one study and those with zero frequency in any of subsets, we analysed the set of GVs with corresponding numbers of individuals, as shown in Table 5.

As seen from Table 5, despite the fact that the number of GVs decreased more than twice after exclusion of nonoverlapping GVs and GVs with zero values, the remaining GVs represent 76–88% of individuals. The data for the set of common GVs with both $n^{h} \neq 0$ and $n^{s} \neq 0$ are displayed in Figure 3.

In order to estimate quantitatively the degree of consistency of the data, we stratified the set of common GVs into four categories using OR, defined by (11):

$$\begin{aligned}
\Omega_{11} &: \left\{\text{OR}^{\text{EIRA}} > 1 \;\&\; \text{OR}^{\text{NARAC}} > 1\right\};\\
\Omega_{12} &: \left\{\text{OR}^{\text{EIRA}} > 1 \;\&\; \text{OR}^{\text{NARAC}} < 1\right\};\\
\Omega_{21} &: \left\{\text{OR}^{\text{EIRA}} < 1 \;\&\; \text{OR}^{\text{NARAC}} > 1\right\};\\
\Omega_{22} &: \left\{\text{OR}^{\text{EIRA}} < 1 \;\&\; \text{OR}^{\text{NARAC}} < 1\right\}.
\end{aligned} \tag{20}$$

The result of this is summarized in Table 6 and only GVs with consistency in direction of the effect (risk or protectivity) were selected for futher analyses; see Tables 7 and 8. At the next step, using (17) for the average OR for each GV and (18) at the level $\alpha = 0.05$, we can construct the plot for the Ω_{11} group of GVs: $\langle\text{OR}\rangle$, $\text{OR}^{\text{tr}}_{\nu,\text{upper}}$ and the "experimental" value of OR. Subsequently, for the Ω_{22} group, where all $\text{OR} < 1$, we used $\text{OR}^{\text{EIRA}}(h/s) = 1/\text{OR}^{\text{EIRA}}_{\Omega_{22}}(s/h)$, $\text{OR}^{\text{NARAC}}(h/s) = 1/\text{OR}^{\text{NARAC}}_{\Omega_{22}}(s/h)$. The results are displayed in Figure 4.

In this comparison, the NARAC group has a much greater size of ORs. However, some of the NARAC GVs happen to be inconsistent with the ones from EIRA. Therefore, only those GVs which exceed the upper 5% threshold *in both* cohorts were chosen for further analysis.

We first analyzed the risk groups, Ω_{11}, represented in the left panels of Figure 4 and found that only 6 GVs, for which

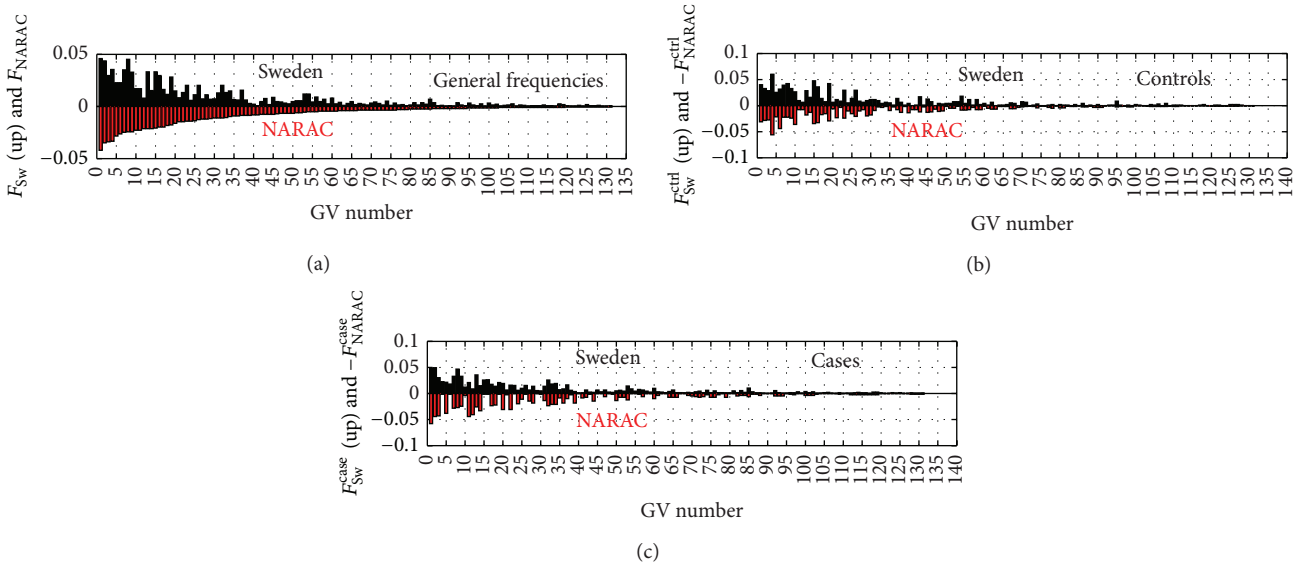

FIGURE 2: Comparison of frequencies of GVs for EIRA (black bars) and NARAC (red bars). All GVs are ordered with respect to general frequencies of NARAC. Frequencies in (a) are defined as $F_{\text{coh}} = (n_H^{(\text{coh})} + n_I^{(\text{coh})})/N_{\text{tot}}^{(\text{coh})}$, coh = EIRA, NARAC.

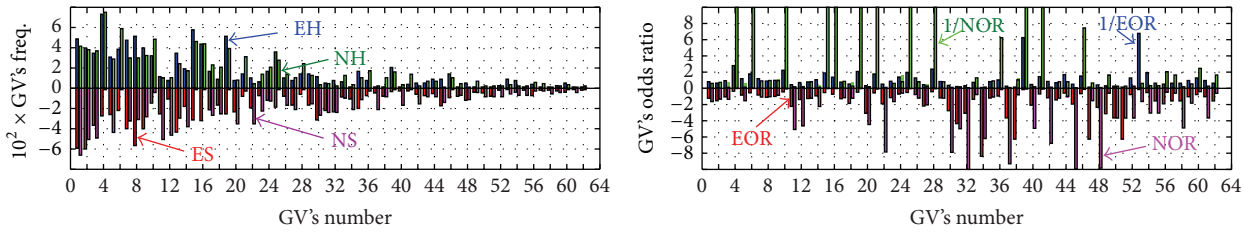

FIGURE 3: Comparison of the data for EIRA and NARAC subsets of GVs that do not contain either $n^h = 0$ or $n^s = 0$. All GVs are sorted with respect to frequencies $f_{\nu,\text{NARAC}}^{\text{tot}} = (n_{\nu,\text{NARAC}}^h + n_{\nu,\text{NARAC}}^s)/\sum_\nu (n_{\nu,\text{NARAC}}^h + n_{\nu,\text{NARAC}}^s)$; (a) displays frequencies of GVs: f_{EIRA}^h (blue), f_{NARAC}^h (green), $-f_{\text{EIRA}}^s$ (red), $-f_{\text{NARAC}}^s$ (lilac); the case-imaging with minus sign is used for a better visibility; EH,ES and NH,NS denote the frequencies of EIRA′ (E) and NARAC′ (N) GVs for healthy (H) controls and cases (S). (b) Odds ratios for GVs, the color notations are the same as for frequencies, and the notations are $1/\text{EOR} \equiv \text{OR}^{\text{EIRA}}(h/s)$, $1/\text{NOR} \equiv \text{OR}^{\text{NARAC}}(h/s)$, $-\text{EOR} \equiv -\text{OR}^{\text{EIRA}}(s/h)$, $-\text{NOR} \equiv -\text{OR}^{\text{NARAC}}(s/h)$. There are several GVs that demonstrate OR > 9.

$\text{OR}_{\nu,\text{exp}} > \text{OR}_{\text{threshold}}^{\text{right}}$ in both cohorts, give consistent results. Let us inspect the probabilities of occasionally finding an OR greater than $\text{OR}_{\nu,\text{exp}}$ in the selected GVs. The result is presented in Figure 5.

Thus, since these subgroups fulfill both criteria, namely, the GVs display OR beyond the fluctuation width near $\langle \text{OR} \rangle$ for chosen $\alpha = 0.05$ and show a low probability of finding $\text{OR}_\nu^{\text{rand}} > \text{OR}_{\nu,\text{exp}}$; the combination of variants corresponding to this subgroup GVs should be considered as most probable candidates for an association with RA.

Among the subgroup of "protective" genotype combinations, which are represented in the right panels of Figure 4, only the first eight GVs give consistent results in both the EIRA and NARAC subgroups. In contrast, the NARAC GVs from 9 to 15, which display $\text{OR}_{\text{exp}} < \text{OR}_{\text{threshold}}^{\text{left}}$, do not fulfil similar requirements in the EIRA subgroup. The values of $-\log_{10} P(\text{OR}_\nu^{\text{rand}} < \text{OR}_{\nu,\text{exp}})$ for the consistent GVs in this subgroup are presented in Figure 6.

The probabilities shown in Figure 6 suggest that a protective role is played by the combinations of variants, presented by these GVs.

In summary, we found that there is a family of GVs in both study populations that follows the same direction of association, although the absolute values of the effects displayed in the NARAC study are more extreme than the ones from the EIRA study for both protective and risk GVs. We chose to use these GVs, with the aim of finding more specific effects from individual genetic variations for further optimization of analyses.

3.3. Candidate Combinations of Genotypes. We used all the available combinations from our genotyping data to differentiate GVs associated with disease. However, the functional consequences from these variations are likely to correspond only to a limited number of selected markers and most of the markers are not important for the association study.

Table 6: Consistent and inconsistent groups of GVs: Ω_{11} group with $N_{11}^{\mathrm{GV}} = 31$ represents GVs with increased risk in both cohorts, whereas Ω_{22} group with $N_{22}^{\mathrm{GV}} = 15$ represents GVs with decreased risk of RA. Ω_{12} and Ω_{21} groups represent inconsistent GVs and were excluded from further analyses.

		N_{11}^t	N_{11}^h	N_{11}^s			N_{12}^t	N_{12}^h	N_{12}^s
$N_{11}^{\mathrm{GV}} = 31$;	EIRA	1192	295	897	$N_{12}^{\mathrm{GV}} = 6$;	EIRA	147	42	105
	NARAC	822	296	526		NARAC	103	63	40
		N_{21}^t	N_{21}^h	N_{21}^s			N_{22}^t	N_{22}^h	N_{22}^s
$N_{21}^{\mathrm{GV}} = 8$;	EIRA	263	106	157	$N_{22}^{\mathrm{GV}} = 15$:	EIRA	670	334	336
	NARAC	205	91	114		NARAC	440	413	27

Table 7: Contrast GVs kk = 11 of individuals in strongest-risk group. Last row displays total number of individuals (the sum of corresponding columns). For briefness the notations are used: $|S_1\rangle = |\mathrm{rs6314}\rangle$, $|S_2\rangle = |\mathrm{rs977003}\rangle$, $|S_3\rangle = |\mathrm{rs1328674}\rangle$, $|S_4\rangle = |\mathrm{rs2070037}\rangle$, $|S_5\rangle = |\mathrm{rs6313}\rangle$, $|S_6\rangle = |\mathrm{rs6311}\rangle$, and $|\Upsilon\rangle = |\mathrm{SE}\rangle$.

Cohorts						Genetic vectors						
EIRA			NARAC			$\lvert S_1\rangle$	$\lvert S_2\rangle$	$\lvert S_3\rangle$	$\lvert S_4\rangle$	$\lvert S_5\rangle$	$\lvert S_6\rangle$	$\lvert\Upsilon\rangle$
$n_{\nu,11}^t$	$n_{\nu,11}^h$	$n_{\nu,11}^s$	$n_{\nu,11}^t$	$n_{\nu,11}^h$	$n_{\nu,11}^s$	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
17	1	16	18	3	15	CC	CA	CC	TT	GG	CC	D
41	6	35	32	7	25	CC	AA	CC	TT	AG	CT	D
41	6	35	105	2	17	CC	AA	CC	CT	AG	CT	D
47	9	38	56	9	36	CC	CA	CC	TT	AG	CT	D
56	9	47	93	3	19	CC	AA	CC	CT	GG	CC	D
37	4	33	30	4	17	CC	CA	CC	CT	AG	CT	D
239	35	204	157	28	129	$\Leftarrow \sum$						

Table 8: Contrast GVs kk = 22 of "most healthy" controls. Last row displays the sum for corresponding columns. For briefness the notations are used: $|S_1\rangle = |\mathrm{rs6314}\rangle$, $|S_2\rangle = |\mathrm{rs977003}\rangle$, $|S_3\rangle = |\mathrm{rs1328674}\rangle$, $|S_4\rangle = |\mathrm{rs2070037}\rangle$, $|S_5\rangle = |\mathrm{rs6313}\rangle$, $|S_6\rangle = |\mathrm{rs6311}\rangle$, and $|\Upsilon\rangle = |\mathrm{SE}\rangle$.

EIRA			NARAC			$\lvert S_1\rangle$	$\lvert S_2\rangle$	$\lvert S_3\rangle$	$\lvert S_4\rangle$	$\lvert S_5\rangle$	$\lvert S_6\rangle$	$\lvert\Upsilon\rangle$
$n_{\nu,22}^t$	$n_{\nu,22}^h$	$n_{\nu,22}^s$	$n_{\nu,22}^t$	$n_{\nu,22}^h$	$n_{\nu,22}^s$	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
92	45	47	41	40	1	CC	AA	CC	TT	AG	CT	N
78	40	38	35	34	1	CC	AA	CC	CT	AG	CT	N
63	30	33	52	51	1	CC	CA	CC	CT	AG	CT	N
98	57	41	66	65	1	CC	CA	CC	TT	AG	CT	N
21	16	5	15	14	1	CT	CA	CC	TT	AG	CT	N
47	25	22	45	42	3	CC	CA	CC	TT	AA	TT	N
56	28	28	25	24	1	CC	AA	CC	TT	GG	CT	N
20	11	9	22	21	1	CC	AA	CC	TT	AA	TT	N
475	252	223	301	291	10	$\Leftarrow \sum$						

Therefore, for further studies it would be desirable to reduce the length of the GVs and to identify the most influential variants within GVs. By excluding unnecessary genetic variants from GVs, we could reconstruct shorter GVs, to facilitate inspection of the genotype content of the GVs in two contrast groups, chosen via two consequent statistical criteria.

In order to compare the results from the contrasting GVs of control and cases with different total numbers of individuals, we sum all IMs (see (4)) in each of categories 11 and 22 with the weights $n_{\nu,kk}^{\alpha}/N_{\alpha}^{\mathrm{sel}}$, where $n_{\nu,kk}^{\alpha}$ is number of individual cases ($\alpha = s$, or "sick") or controls ($\alpha = h$, or "healthy") in GV ν, belonging to the class $kk = 11, 22$; here $N_{\alpha}^{\mathrm{sel}}$ are the numbers of individuals in the selected group kk, $N_{\alpha,kk}^{\mathrm{sel}} = \sum_{\nu\in\Omega_{kk}^{\alpha}} n_{\nu,kk}^{\alpha}$; notice that here the normalization is different from the one used in (7):

$$N_{i\gamma,kk}^{\alpha} = \frac{1}{N_{\alpha,kk}^{\mathrm{sel}}} \sum_{\nu\in\Omega_{kk}^{\alpha}} n_{\nu,kk}^{\alpha} \left\langle \Psi^{\alpha} \right| X_{\nu i}^{\gamma} \left| \Psi^{\alpha} \right\rangle, \qquad \alpha = h, s; \quad kk = 11, 22. \tag{21}$$

As expected, our data demonstrate strong association of RA with SE alleles (see Table 9). At the same time, the SNP rs1328674 does not show an association with RA, in contrast to a previous report [9], while we detected an association

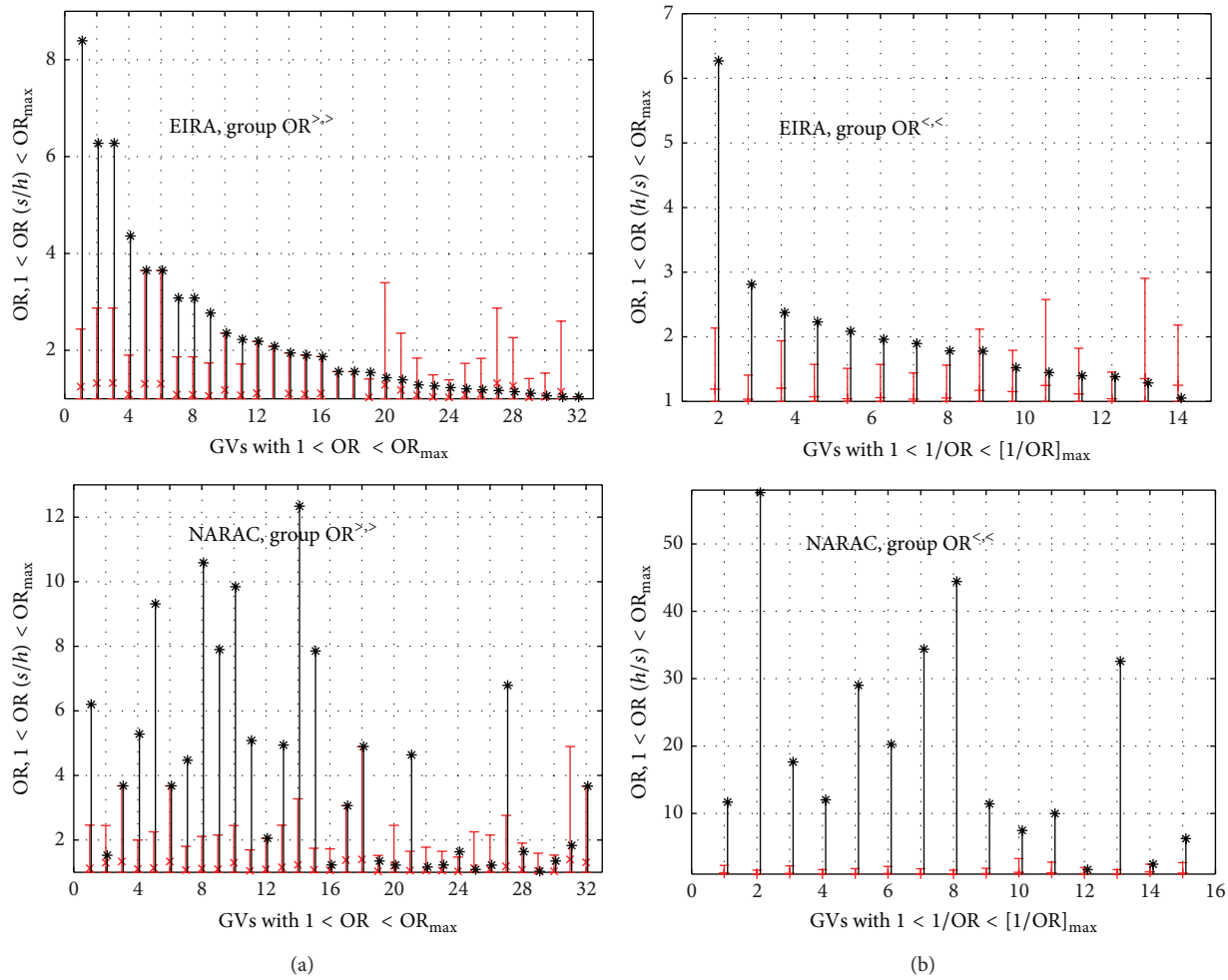

FIGURE 4: Odds ratios for EIRA and NARAC for common GVs "without zeros" ($n_s \neq 0$ and $n_h \neq 0$). The red stars represent the expectation value ⟨OR⟩ of random OR (notice that ⟨OR⟩ is slightly different from unity due to asymmetry and discreteness of the distribution for OR); error bars show the interval that is hit by OR with randomly chosen n_ν^s from the interval $(0, 1, 2, \ldots, n_\nu^t - 1)$ with 95% probability; black stars display observed OR for given GV. Thus, if $OR_{\nu,exp}$ is above the upper error bar, the null hypothesis that the observed value of OR can arise occasionally is not supported with 5% threshold value (for the random variable $n_s = 0, 1, 2, \ldots, n_t - 1$).

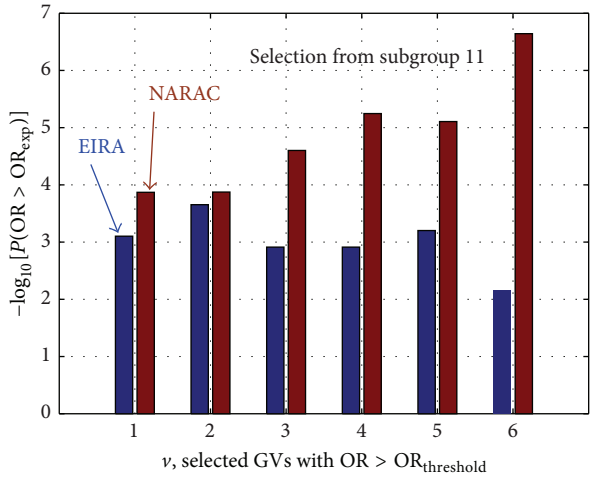

FIGURE 5: Negative logarithm of the probability to find $OR_\nu^{rand} > OR_{\nu,exp}$ in the subgroup Ω_{11} : $\{OR^{EIRA} > 1 \ \& \ OR^{NARAC} > 1\}$.

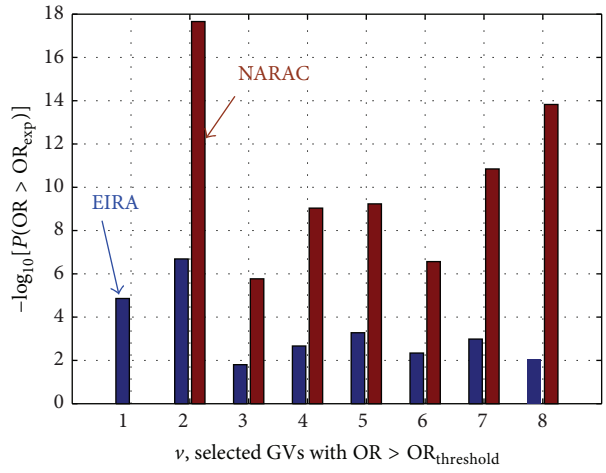

FIGURE 6: The values of $-\log_{10} P(OR_\nu^{rand} < OR_{\nu,exp})$ for the consistent GVs in the subgroup Ω_{22} : $\{OR^{EIRA} < 1 \ \& \ OR^{NARAC} < 1\}$.

Table 9: Averaged index matrices of contrast EIRA GVs: 11 = cases are followed by "#," 22 = controls are followed by "∗," and their difference is by "∗∗." Strongest changes are marked by bold fonts. For briefness the notations are used: $|S_1\rangle$ = |rs6314⟩, $|S_2\rangle$ = |rs977003⟩, $|S_3\rangle$ = |rs1328674⟩, $|S_4\rangle$ = |rs2070037⟩, $|S_5\rangle$ = |rs6313⟩, $|S_6\rangle$ = |rs6311⟩, and $|Y\rangle$ = |SE⟩.

EIRA	$\|S_1\rangle$	$\|S_2\rangle$	$\|S_3\rangle$	$\|S_4\rangle$	$\|S_5\rangle$	$\|S_6\rangle$	$\|Y\rangle$
	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
	CC	AA	CC	CC	AA	CC	No
$\gamma_{i,s} = 1$	$1^{\#}$	$0.5735^{\#}$	$1^{\#}$	$0^{\#}$	$0^{\#}$	$0.3088^{\#}$	$0^{\#}$
$\gamma_{i,h} = 1$	0.9365^{*}	0.4921^{*}	1^{*}	0^{*}	0.1428^{*}	0.1111^{*}	1^{*}
$\gamma_{i,d} = 1$	0.0635^{**}	0.0815^{**}	0^{**}	0^{**}	$\mathbf{-0.1429}^{**}$	$\mathbf{0.1977}^{**}$	-1^{**}
	CT	AC	CT	CT	AG	CT	Yes
$\gamma_{i,s} = 2$	$0^{\#}$	$0.4265^{\#}$	$0^{\#}$	$0.5637^{\#}$	$0.6912^{\#}$	$0.6912^{\#}$	$0^{\#}$
$\gamma_{i,h} = 2$	0.0635^{*}	0.5079^{*}	0^{*}	0.2778^{*}	0.7460^{*}	0.7460^{*}	0^{*}
$\gamma_{i,d} = 2$	-0.0635^{**}	-0.0815^{**}	0^{**}	$\mathbf{0.2859}^{**}$	-0.0549^{**}	-0.0549^{**}	0^{**}
	TT	CC	TT	TT	GG	DD	Double
$\gamma_{i,s} = 3$	$0^{\#}$	$0^{\#}$	$0^{\#}$	$0.4363^{\#}$	$0.3088^{\#}$	$0^{\#}$	$1^{\#}$
$\gamma_{i,h} = 3$	0^{*}	0^{*}	0^{*}	0.7222^{*}	0.1111^{*}	0.1429^{*}	0^{*}
$\gamma_{i,d} = 3$	0^{**}	0^{**}	0^{**}	$\mathbf{-0.2859}^{**}$	$\mathbf{0.1977}^{**}$	$\mathbf{-0.1429}^{**}$	1^{**}

Table 10: Averaged index martices of contrast NARAC GVs: 11 = "sick" are followed by "#," 22 = "healthy" are followed by "∗," and their difference is followed by "∗∗." Strongest changes are marked by bold fonts. For briefness the notations are used: $|S_1\rangle$ = |rs6314⟩, $|S_2\rangle$ = |rs977003⟩, $|S_3\rangle$ = |rs1328674⟩, $|S_4\rangle$ = |rs2070037⟩, $|S_5\rangle$ = |rs6313⟩, $|S_6\rangle$ = |rs6311⟩, and $|Y\rangle$ = |SE⟩.

NARAC	$\|S_1\rangle$	$\|S_2\rangle$	$\|S_3\rangle$	$\|S_4\rangle$	$\|S_5\rangle$	$\|S_6\rangle$	$\|SE\rangle$
	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
	CC	AA	CC	CC	AA	CC	No
$\gamma_{i,s} = 1$	$1^{\#}$	$0.4729^{\#}$	$1^{\#}$	$0^{\#}$	$0^{\#}$	$0.2636^{\#}$	$0^{\#}$
$\gamma_{i,h} = 1$	0.9519^{*}	0.4089^{*}	1^{*}	0^{*}	0.2165^{*}	0.0825^{*}	1^{*}
$\gamma_{i,d} = 1$	0.0481^{**}	0.0639^{**}	0^{**}	0^{**}	$\mathbf{-0.2165}^{**}$	$\mathbf{0.1811}^{**}$	-1^{**}
	CT	AC	CT	CT	AG	CT	Yes
$\gamma_{i,s} = 2$	$0^{\#}$	$0.5271^{\#}$	$0^{\#}$	$0.4109^{\#}$	$0.7364^{\#}$	$0.7364^{\#}$	$0^{\#}$
$\gamma_{i,h} = 2$	0.0481^{*}	0.5911^{*}	0^{*}	0.2921^{*}	0.7010^{*}	0.7010^{*}	0^{*}
$\gamma_{i,d} = 2$	-0.0481^{**}	-0.0639^{**}	0^{**}	$\mathbf{0.1188}^{**}$	0.0354^{**}	0.0354^{**}	0^{**}
	TT	CC	TT	TT	GG	TT	Double
$\gamma_{i,s} = 3$	$0^{\#}$	$0^{\#}$	$0^{\#}$	$0.5891^{\#}$	$0.2636^{\#}$	$0^{\#}$	$1^{\#}$
$\gamma_{i,h} = 3$	0^{*}	0^{*}	0^{*}	0.7079^{*}	0.0825^{*}	0.2165^{*}	0^{*}
$\gamma_{i,d} = 3$	0^{**}	0^{**}	0^{**}	$\mathbf{-0.1188}^{**}$	$\mathbf{0.1811}^{**}$	$\mathbf{-0.2165}^{**}$	1^{**}

of the SNPs rs6314, rs977003, rs2070037, rs6313, and rs6311 in the EIRA study. Since not all of these results agree with previously published evaluation at this locus, we compare it with independent data from NARAC; see Table 10.

As can be seen, the data for NARAC fully confirm the conclusions derived from the EIRA data. The numbers for NARAC in Table 10 differ from those for EIRA (see Table 9) only slightly; therefore, the tendencies revealed in both cohorts from the same sets of GVs can be considered to be cross-validated.

Thus, selection of shorter GVs for statistical evaluation demonstrated how the overall analysis could be optimized: $|rs2070037\rangle$, $|rs6313\rangle$, $|rs6311\rangle$, and $|SE\rangle$ remain indicative for the difference between groups, whereas $|rs6314\rangle$, $|rs977003\rangle$, and $|rs1328674\rangle$ have no influence on association in the GVs of both EIRA and NARAC. These cropped GVs can be especially useful for optimizing the analysis of long GVs, since it can improve the statistical confidence of the results.

We identified significant heterogeneity of association of combinations of HTR2A variants, which depends on the absence or presence of a shared epitope allele in the GV. Several interesting observations have been found from application of the method to two RA study populations, EIRA and NARAC.

We found that the number of GVs that describe the study population for the chosen length of GV is relatively small: there are only 161 GVs required for description of the EIRA, with 2767 individuals, and 163 GVs for NARAC, consisting of 1974 people. A very high overlap between Swedish and North American study populations was detected in our analysis: *131 of GVs are common to both study populations*.

Interestingly, some of the GVs contain only healthy controls and do not contain the case counterpart (subset of

solely healthy persons), and *vice versa*; some of the GVs contain only cases and do not contain controls (solely ill). The statistical weight of these GVs, however, appeared to be small. Nevertheless, we have found similar behaviour in two independent cohorts and this observation deserves further investigation. Even after removing these "zero" GVs from consideration, the remaining set of common EIRA and NARAC GVs contains about 80% of the total number of individuals in each cohort.

Our literature search did not give much information about similar methods being used in genetic studies. The closest approach, haplotype analysis, is based on the classical genetic idea of transmission of the marker from parents to offspring and therefore the data in this analysis are arranged around the combination of genetic markers within a single chromosome. In an association study this approach may well serve in allelic mode, but it totally ignores the possible involvement of a second allele. In functional studies, selection of individuals by haplotypes is straightforward for homozygotic states but introduces multiple combinations in the case of heterozygosity and generates an extreme number of genotype groups in comparison with GVA. Experimental detection of haplotypes is a difficult procedure and the common statistical approach for assigning of haplotypes is never 100% unambiguous.

3.4. Genetic Vectors with Zero Values in One of the Subgroups, Experimental Data. In this section we illustrate previously defined "zero GV" detected in the EIRA and NARAC studies of RA. The number of individuals involved in each step of selection is decreased as shown in Table 11.

Thus, of 131 GVs common for EIRA and NARAC, 19 GVs happen to be zero GVs, that is, either $n^s_{\nu,\alpha} = 0$ and $n^h_{\nu,\alpha} \neq 0$ or, *vice versa*, $n^s_{\nu,\alpha} \neq 0$ and $n^h_{\nu,\alpha} = 0$. They contain only about 2% of individuals from the study populations (62/2727 in EIRA and 43/1938 in NARAC). It is interesting to compare the contents of GVs from the "solely sick" (SSGV) and the "solely healthy" (SHGV) subgroups. Since the number of individuals belonging to the classes of interest is small, it would be nice to have a look at the pooled data. The inspection, however, shows that the part of the GVs that belong to the zero GV subgroup in, say, EIRA, often does not belong to this subgroup in NARAC, and *vice versa*. These GVs lose their status of zero GVs in the pooled data.

The insufficient number of individuals in SSGV and SHGV subgroups does not make it possible to draw statistically confident conclusions if we work with each study population separately. We can, however, investigate whether there exists some cross-validated tendency, that is, the one observed in both EIRA and NARAC.

This can be visualized via GV frequency weighted index matrices for zero GVs. Since we are interested only in the groups of zero GVs, we will perform averaging over these subgroups only. In other words, (13) should be modified as follows:

$$\overline{C^{\mathrm{sh}}} = \frac{1}{N_{\mathrm{sh}}} \sum_{\nu \subset \Omega_0^{\mathrm{sh}}} n_\nu^{\mathrm{sh}} C_\nu^{\mathrm{sh}},$$

Table 11: Numbers of GVs and individuals in full cohorts, in common subsets of GVs, and in subsets of "solely healthy" (SH, $n^s_\nu = 0$, $n^h_\nu = n^t_\nu$) and "solely sick" (SS, $n^s_\nu = n^t_\nu$, $n^h_\nu = 0$) individuals. The notations in the table are $\xi_t = \mathbf{N_t^{GV}}/N_t$, that is, (number of GVs)/(number of individuals); $\xi_{\mathrm{PS}} = \mathbf{N_{SS}^{GV}}/N_{\mathrm{SS}}$, that is, (number of SS GVs)/(number of SS individuals); $\xi_{\mathrm{SH}} = \mathbf{N_{SH}^{GV}}/N_{\mathrm{SH}}$, that is, (number of SH GVs)/(number of SH individuals); $N_s^{\mathrm{tot}}/N_h^{\mathrm{tot}}$ is (total number of cases)/(total number of controls); $\xi_t^c = \mathbf{N_{t,common}^{GV}}/N_t^c$, N_s^c/N_h^c is (common number of cases)/(common number of controls); $\xi_{\mathrm{SS}}^{\mathrm{cr}} = \mathbf{N_{SS,cr}^{GV}}/N_{\mathrm{SS}}^{\mathrm{cr}}$, that is, (number of common SS GVs)/(number of common SS individuals); $\xi_{\mathrm{SH}}^{\mathrm{cr}} = \mathbf{N_{SH,cr}^{GV}}/N_{\mathrm{SH}}^{\mathrm{cr}}$, that is, (number of common SH GVs)/(number of common SH individuals).

		EIRA	NARAC
Full cohorts	ξ_t	**161**/2767	**163**/1974
	$N_s^{\mathrm{tot}}/N_h^{\mathrm{tot}}$	1820/947	813/1161
	ξ_{SS}	**52**/152	**28**/61
	ξ_{SH}	**15**/21	**62**/263
GVs	ξ_t^c	**131**/2727	**131**/1938
GVs common for EIRA and NARAC	N_s^c/N_h^c	1789/938	804/1134
	$\xi_{\mathrm{SS}}^{\mathrm{cr}}$	**12**/55	**12**/31
	$\xi_{\mathrm{SH}}^{\mathrm{cr}}$	**7**/13	**7**/12

$$\overline{C^{\mathrm{ss}}} = \frac{1}{N_{\mathrm{ss}}} \sum_{\nu \subset \Omega_0^{\mathrm{ss}}} n_\nu^{\mathrm{ss}} C_\nu^{\mathrm{ss}}, \tag{22}$$

where n_ν^{sh} and n_ν^{ss} are numbers of individuals in the GV ν belonging to the group SH or SS correspondingly; indices sh and ss mean "solely sick" and "solely healthy"; the numbers

$$N_{\mathrm{sh}} = \sum_{\nu \subset \Omega_0^{\mathrm{sh}}} n_\nu^{\mathrm{sh}}, \qquad N_{\mathrm{ss}} = \sum_{\nu \subset \Omega_0^{\mathrm{ss}}} n_\nu^{\mathrm{ss}} \tag{23}$$

are total numbers of individuals in the SSGV and SHGV subgroups and C_ν^{sh} and C_ν^{ss} are index matrices of the GV ν from SH and SS subgroups. Then we can compare EIRA and NARAC results for case-control differences of index matrices, $\overline{C^{\mathrm{ss}}} - \overline{C^{\mathrm{sh}}}$ for zero GVs. The result is shown in Figure 7.

Full coherence in EIRA and NARAC data display is only SE (no SE is protective, while double SE strongly associates with RA, which is expected) and the SNP rs1328674. The positive value of the difference for frequencies of the genotype (rs1328674,CT) means the tendency to disease, whereas the negative one for the genotype (rs1328674,CC) indicates the tendency to protect. All other genotypes do not display consistency between the results for EIRA and NARAC. Thus, although the fact of zero GVs existence is interesting, we have to admit that, due to the absence of (i) the consistency for two study populations and (ii) the possibility of drawing statistically confident conclusions, analysis of zero GVs should be taken with caution when the numbers of observations are too low. In practice, it will need replication study in very big cohorts.

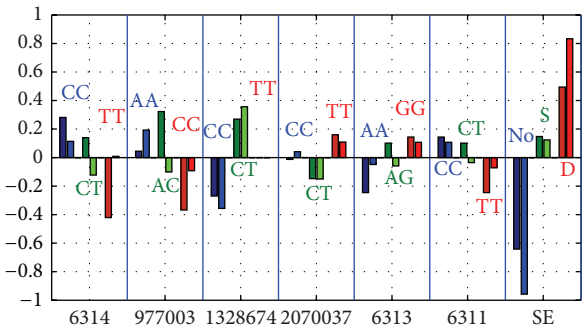

FIGURE 7: EIRA-versus-NARAC comparison charts for selected zero GVs: the differences in genotype frequencies $f^s_{iy} - f^h_{iy}$ for GVs of "solely sick" and "solely healthy" individuals. Positive values indicate a propensity for disease, whereas the negative values suggest that a protective role is played by the corresponding genotype. In each genotype couple, the darker color on the left corresponds to EIRA, while the lighter color on the right presents the result for the NARAC difference of frequencies. S: single SE; D: double SE allele.

4. Conclusions

4.1. On GVA. Our approach, GVA, is the method for genetic analysis based on testing of association to a complex trait simultaneously for combinations of variants (multiple markers) *inherent to individuals* without any assumptions about their statistical relations. It is based on ascribing to each person a multiple-spin quantum state, GV, with subsequent description of the study population as an ensemble of spin chains with a given (determined by experimental data) set of allowed states. The full data set is sorted into subsets, each of which is described by a single GV and the numbers of cases and controls representing this GV. We successfully tested this approach for the set of experimental data for RA and validated the previously known association with shared epitope alleles.

GVA has several advantages:

First, since GVA works with the genotype combinations as a single entity, without decoupling of random variables, the issue of possible statistical independence (as well as interaction, influence of the linkage disequilibrium, etc.) of separate genotypes does not affect analyses.

Second, being formulated for genotypes, it is free of the uncertainties connected with the use of an allelic model or haplotypes: it is, therefore, not necessary to guess a mode of dominance. At the same time, the frequencies of separate genotypes, or their pairs, triples, and so forth, can be easily expressed in terms of GV variables, which means that other analyses could be easily performed by simplification of the GV algorithm.

Third, from the perspective of functional genetics and personalized medicine, this approach provides the opportunity to directly assign individuals belonging to the risk group of interest to a genetically determined subset defined by GV. Since GVA includes in the analysis only the genotype (phenotype) combinations which are met in the population study, the number of combinations that need to be analyzed is significantly decreased compared to the number of random combinations of genetic markers; therefore, the volume of necessary calculations is greatly reduced. The problem of multiple combinations of haplotypes is also resolved in GVA.

The probability distribution for OR of each GV, which we derived from the hypergeometric distribution for the random variable "number of cases in GV," happens to be asymmetric and is not Gaussian. For this reason, the standard method for evaluating the confidence accuracy could not be used. Thus, we instead evaluated the thresholds of the OR interval where OR fluctuations exceed 5% for each OR separately. The statistics of each GV depends on three parameters: the total number of cases, the total number of controls in the study population, and the total number of individuals (cases plus controls), described by the GV in question. Only those of the GVs for which OR falls outside of the interval of large OR fluctuations have been selected for further analysis. The dependence of the OR statistics for a particular GV on the number of individuals "belonging" to this GV creates the obvious disadvantage of GVA: with an increase of the GV length (the number of SNPs included in the GV) the total number of GVs describing the study population approaches the number of individuals, since each person has unique DNA and, therefore, statistical analysis becomes invalid.

Regarding optimization of GVA and reducing the number of markers in GV (cropping the vector size), we suggested using index matrices, uniquely characterizing each GV. By using this approach a comparative analysis of the "genotype content" can be done for the sets of GVs of interest, identifying the sets of genotypes that influence the difference between the control and case groups.

It is worth noting that GVA does not require any assumptions about statistical coupling between variables. The study of interaction and synergetic effects between particular SNPs can also be performed in terms of GVA. Indeed, the answer to the question whether pairs, or triples, of some of SNPs produce a much stronger association to the phenotype of interest than is expected from independent factors can be obtained from the analysis of the correlation functions between SNPs. Examples of this type of analysis have been discussed in [2].

4.2. On RA in EIRA and NARAC. Our analysis reveals that the majority of GV groups, 76.7% (or 81.4% of individuals in all GVs, Table 6), are consistent for the effect between EIRA and NARAC. The degree of inconsistence could be possibly explained by low number of observation and randomness, but also by influence of different environmental factors or by different genetic architecture of the locus (between included SNPs and at flanking regions).

We confirm the well-known fact that shared epitope (SE) is strongly associated with RA. What seems to be significant here is that the whole set of genotypes displays quite large differences in frequencies in the presence and absence of SE. The SNPs rs2070037, rs6313, and rs6311 seem to play a more important role in RA formation than they were considered to do in a previous study [9].

List of Abbreviations

EIRA: Swedish Epidemiological Investigation of Rheumatoid Arthritis
GV: Genetic vector
GVA: Genetic vector approach
NARAC: North American Rheumatoid Arthritis Consortium
OR: Odds ratio
RA: Rheumatoid arthritis
SNP: Single-nucleotide polymorphism.

Conflict of Interests

The authors declare that they have no conflict of interests.

Authors' Contribution

Igor Sandalov and Leonid Padyukov are responsible for design of the study; Igor Sandalov is responsible for computations and model evaluation; Igor Sandalov and Leonid Padyukov are responsible for preparation of the paper.

Acknowledgments

This study was supported by the Swedish Research Council, Vinnova (COMBINE Grant), and Innovative Medicines Initiative Joint Undertaking, 115142-2 BTCURE. The authors would like to thank Janet Ahlberg for help with English editing of the paper.

References

[1] K. van Steen, "Travelling the world of gene-gene interactions," *Briefings in Bioinformatics*, vol. 13, no. 1, pp. 1–19, 2012.

[2] I. Sandalov and L. Padyukov, "Genetic vectors approach in a study of fine structure of interaction between risk haplotype of HTR2A and HLA-DRB1 shared epitope alleles in rheumatoid arthritis," in *Between the Lines of Genetic Code. Genetic Interactions in Understanding Disease and Complex Phenotypes*, L. Padyukov, Ed., pp. 151–174, Elsevier, 2014.

[3] N. Chatterjee, Y.-H. Chen, S. Luo, and R. J. Carroll, "Analysis of case-control association studies: SNPs, imputation and haplotypes," *Statistical Science*, vol. 24, no. 4, pp. 489–502, 2009.

[4] D. Jawaheer, R. F. Lum, C. I. Amos, P. K. Gregersen, and L. A. Criswell, "Clustering of disease features within 512 multicase rheumatoid arthritis families," *Arthritis and Rheumatism*, vol. 50, no. 3, pp. 736–741, 2004.

[5] L. Klareskog, P. Stolt, K. Lundberg et al., "A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination," *Arthritis and Rheumatism*, vol. 54, no. 1, pp. 38–46, 2006.

[6] F. C. Arnett, S. M. Edworthy, D. A. Bloch et al., "The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis," *Arthritis and Rheumatism*, vol. 31, no. 3, pp. 315–324, 1988.

[7] D. Jawaheer, M. F. Seldin, C. I. Amos et al., "A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases," *The American Journal of Human Genetics*, vol. 68, no. 4, pp. 927–936, 2001.

[8] E. Lundström, H. Källberg, M. Smolnikova et al., "Opposing effects of HLA-DRB1*13 alleles on the risk of developing anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis," *Arthritis and Rheumatism*, vol. 60, no. 4, pp. 924–930, 2009.

[9] A. Kling, M. Seddighzadeh, L. Ärlestig, L. Alfredsson, S. Rantapää-Dahlqvist, and L. Padyukov, "Genetic variations in the serotonin 5-HT2A receptor gene (HTR2A) are associated with rheumatoid arthritis," *Annals of the Rheumatic Diseases*, vol. 67, no. 8, pp. 1111–1115, 2008.

[10] D. Karevski, *Ising Quantum Chains. Statistical Mechanics*, Université Henri Poincaré Nancy 1, Nancy, France, 2005.