genes

MDPI

Article

Gene-Based Nonparametric Testing of Interactions Using Distance Correlation Coefficient in Case-Control Association Studies

Yingjie Guo [1,2,*], Chenxi Wu [3], Maozu Guo [1,4,*], Xiaoyan Liu [1] and Alon Keinan [2]

1 School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China; liuxiaoyan@hit.edu.cn
2 Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA; alon.keinan@cornell.edu
3 Department of Mathematics, Rutgers University, Piscataway, NJ 08854, USA; wuchenxi2013@gmail.com
4 Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
* Correspondence: yjguo0625@gmail.com (Y.G.); guomaozu@bucea.edu.cn (M.G.)

Received: 6 November 2018; Accepted: 27 November 2018; Published: 5 December 2018

Abstract: Among the various statistical methods for identifying gene–gene interactions in qualitative genome-wide association studies (GWAS), gene-based methods have recently grown in popularity because they confer advantages in both statistical power and biological interpretability. However, most of these methods make strong assumptions about the form of the relationship between traits and single-nucleotide polymorphisms, which result in limited statistical power. In this paper, we propose a gene-based method based on the distance correlation coefficient called gene-based gene-gene interaction via distance correlation coefficient (GBDcor). The distance correlation (dCor) is a measurement of the dependency between two random vectors with arbitrary, and not necessarily equal, dimensions. We used the difference in dCor in case and control datasets as an indicator of gene–gene interaction, which was based on the assumption that the joint distribution of two genes in case subjects and in control subjects should not be significantly different if the two genes do not interact. We designed a permutation-based statistical test to evaluate the difference between dCor in cases and controls for a pair of genes, and we provided the *p*-value for the statistic to represent the significance of the interaction between the two genes. In experiments with both simulated and real-world data, our method outperformed previous approaches in detecting interactions accurately.

Keywords: genome-wide association studies; qualitative trait; gene–gene interaction; distance correlation coefficient

1. Introduction

Genome-wide association studies (GWAS) are a well-established and effective method of identifying genetic loci associated with common diseases or traits, and they have identified over 65,000 unique single-nucleotide polymorphisms (SNPs) that are associated with various traits or diseases [1–5]. Earlier GWAS analysis strategies were based largely on single-locus models, which tested the association between individual markers and a given phenotype independently. Although this type of approach has identified many regions of disease susceptibility successfully, most of the SNPs that have been identified have small effect sizes that failed to account fully for the heritability of complex traits. Genetic interaction has been hypothesized to play an important role in the genetic basis of complex diseases and traits, [6,7], and it has been one of the possible solutions to the problem of "missing heritability" [8–10].

Even if genetic interaction explains only a tiny fraction of "missing heritability", it can still provide some biological insight into the pathway by aiding the construction of novel gene pathway topologies.

The first investigations on genetic interactions were at the SNP level, in which various statistical methods, which included logistic regression [11–13], odds-ratio [14], linkage disequilibrium (LD), [15–17], and entropy-based statistics [18,19], were employed to detect SNP–SNP interactions (i.e., epistasis). Other techniques that have been used to study SNP–SNP interactions include multifactor dimensionality reduction (MDR) [20], Tuned RelieF (TuRF) [21], Bayesian epistasis association mapping (BEAM) [6], Tree-based epistasis association mapping (TEAM) [22], Boolean operation-based screening and testing (BOOST) [23], and permutation-based Random Forest (pRF) [24]. These marker-based methods have encountered some common challenges, such as the complexity that arises from the large number of pairwise or higher-order tests because all pairs or groups of SNPs have to be considered and because of the extensive multiple testing correction, which weakens their statistical power. In this paper, we aim to improve the power of the detection of gene–gene interactions by moving beyond the SNP level and, instead, consider all potential pairs of SNPs from each of a pair of genes in a single, gene-based, interaction detection.

In the study of the main (marginal) associations in GWAS, gene-based approaches have been successful, and therefore, it might be worth extending it to the analysis of interaction between genes. There are several potential advantages of this approach. First, it can reduce the number of pairwise tests substantially, because there are usually many fewer genes than SNPs. For example, detection of pairwise, gene-based interactions for 20,000 genes requires $\sim 2 \times 10^8$ tests, but for three million SNPs, the marker-based interaction tests require more than $\sim 5 \times 10^{12}$. Second, gene-based approaches might have greater statistical power, because a gene contains more information than a single SNP and because there might be multiple ways for genes to interact with each other that are aggregated; this is also true when doing a gene-based study for main effects [25,26]. Third, it might be easier to incorporate prior biological knowledge with this approach (e.g., information on protein–protein interactions (PPI) or known membership of genes in pathways). Finally, the results of gene-based analysis may have more meaningful biological implications and be more interpretable.

In the work of Peng et al. [27], canonical correlation between two genes was performed on both the case and the control groups. A U-statistic called canonical correlation-based U statistic (CCU) was used to measure the difference in the correlation between these two genes, which was used to indicate the presence of interaction. A limitation of this method was that in the correlation analysis, only linear relations were considered. To overcome this limitation, Yuan et al. [28] and Larson et al. [29] extended CCU to kernelized CCU (KCCU), where the canonical correlation analysis was kernelized to account for possible non-linearity. Li et al. [30] introduced another method called the gene-based information gain method (GBIGM), which was entropy-based and non-parametric. More recently, Emily [31] developed a new method called gene-based gene-gene interaction test(AGGrEGATOr), which combined the p-values in marker-level interaction tests to measure the interaction between two genes. Earlier, this strategy was used successfully by Ma et al. [32] for the detection of interaction for quantitative phenotypes.

In this paper, we propose a novel method to identify gene-level, gene–gene interactions among case control studies of complex phenotypes based on the distance correlation coefficient called gene-based gene-gene interaction via distance correlation coefficient (GBDcor). Distance correlation [33–35] quantifies all types of dependent relationships between two random vectors with arbitrary, but not necessarily equal, dimensions, which is better than Pearson's correlation, which only focuses on the linear relationship. Distance correlation has already been used in bioinformatics to detect co-expression genes [36] and imaging genetics associations [37]. We use the difference in dependence relationships between case samples and control samples as an indicator of gene–gene interaction, which is based on the assumption that the joint distribution of two genes in case subjects and in control subjects should not be significantly different if the two genes do not interact (i.e., independent) under the case-control status. Experiments on semi-empirical data showed that the distance correlation with

permutation strategy yielded better power to detect underlying gene-based gene–gene interactions in a large range of settings, and the application to real datasets verified that GBDcor identified gene–gene interactions accurately.

2. Materials and Methods

In this section, we detail the statistical procedure for GBDcor. We then present the various settings for semi-empirical simulation studies for the type-I error rate and for the power to detect gene–gene interaction. Finally, we describe a real rheumatoid arthritis dataset from the wellcome trust case control consortium (WTCCC) database to evaluate our method in a real situation.

2.1. GBDcor

2.1.1. Preliminaries and Notation

Suppose that we have random samples:

$$(\boldsymbol{G}_{1,i}, \boldsymbol{G}_{2,i}) \in \mathcal{R}^{p+q}, i = 1, 2, ..., n$$

where:

$$\boldsymbol{G}_{1,i} = (g_{1,i,1}, g_{1,i,2}, ..., g_{1,i,p}), \boldsymbol{G}_{2,i} = (g_{2,i,1}, g_{2,i,2}, ..., g_{2,i,q}), i = 1, 2, ..., n$$

where $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ represent genes with p and q SNPs, respectively. $g_{k,i,j} \in \{0, 1, 2\}$ is the number of copies of the minor allele for SNP j in gene k of sample i. We focus on the case-control data that $y_i \in \{0, 1\}$ is a categorical label, where 1 represents case subjects and 0 represents control subjects. Here, $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ are assumed to take values in $\{0, 1, 2\}^p$ and $\{0, 1, 2\}^q$, respectively, where $(\boldsymbol{G}_{1,i}, \boldsymbol{G}_{2,i}) \in \{0, 1, 2\}^{p+q}, i = 1, 2, ..., n$, is a random sample from the joint distribution of $(\boldsymbol{G}_1, \boldsymbol{G}_2)$.

In this work, to investigate whether there is a statistical interaction between two genes in a qualitative phenotype, we combine the distance correlation with the permutation strategy to test whether two genes interact. Our approach is based on the intuition that, if there is no interaction between two genes, then, if they are independent of the case set, then they should be independent of the control set; if they are dependent on the case set, they should be dependent on the control set also, and the "strength" of such dependence should be the same on the case and control set. The degree of dependence between two random variables can be measured by Pearson's correlation coefficients. However, it can only measure linear dependency and not nonlinear dependency, and it is not very convenient for random variables that take a value in $\mathcal{R}^n$; hence, we propose measuring them by the distance correlation coefficient instead.

2.1.2. Distance Correlation

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random variables in $\mathcal{R}^n$ with finite first moments, then their distance covariance, denoted by $dCov(\boldsymbol{X}, \boldsymbol{Y})$, and distance correlation coefficients, denoted by $R^2(\boldsymbol{X}, \boldsymbol{Y})$, are defined in ([33]). They satisfy the following properties:

- $\mathcal{R}(\boldsymbol{X}, \boldsymbol{Y})$ is defined for $\boldsymbol{X}$ and $\boldsymbol{Y}$ in arbitrary dimensions.
- $\mathcal{R}^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{dCov^2(\boldsymbol{X}, \boldsymbol{Y})}{\sqrt{dCov^2(\boldsymbol{X}, \boldsymbol{X}) dCov^2(\boldsymbol{Y}, \boldsymbol{Y})}}$
- $\mathcal{R}(\boldsymbol{X}, \boldsymbol{Y}) = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent.
- $0 \leq \mathcal{R}(\boldsymbol{X}, \boldsymbol{Y}) \leq 1$

The proofs can be found in [33]. In particular, Property 4 above tells us that the distance correlation can be used to measure the degree of dependency between two random variables.

If there are n samples $(X_i, Y_i), i = 1, ..., n$, according to [33], the distance covariance and distance correlation can be estimated by the sample distance covariance and sample distance correlation, which we will describe below.

Let $A_{i,j}, B_{i,j}$ be the centered distance matrix of the samples X_i, Y_i. In other words,

$$A_{i,j} = |X_i - X_j| - \frac{1}{n}\sum_k |X_k - X_j| - \frac{1}{n}\sum_l |X_i - X_l| + \frac{1}{n^2}\sum_{k,l} |X_k - X_l| \tag{1}$$

$$B_{i,j} = |Y_i - Y_j| - \frac{1}{n}\sum_k |Y_k - Y_j| - \frac{1}{n}\sum_l |Y_i - Y_l| + \frac{1}{n^2}\sum_{k,l} |Y_k - Y_l| \tag{2}$$

Then, the sample distance covariance is defined as:

$$dCov_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{i,j} A_{i,j}B_{i,j} \tag{3}$$

and the sample distance correlation coefficient is:

$$\mathcal{R}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{dCov_n^2(\boldsymbol{X}, \boldsymbol{Y})}{\sqrt{dCov_n^2(\boldsymbol{X}, \boldsymbol{X}) dCov_n^2(\boldsymbol{Y}, \boldsymbol{Y})}} \tag{4}$$

2.1.3. dCor with Permutation Strategy

Assume there are n_1 cases and n_2 controls in a case-control study for a pair of genes $\boldsymbol{G_1}$ with p SNPs and $\boldsymbol{G_2}$ with q SNPs. Let $dCor_n = \mathcal{R}_n^2(\mathbf{G_1}, \mathbf{G_2})$ be the sample distance correlation between Gene 1 and Gene 2 for a subsample of size n. First, we calculate the $dCor_{n_1}^C = \mathcal{R}_{n_1}^2(\mathbf{G_1^C}, \mathbf{G_2^C})$, $dCor_{n_2}^D = \mathcal{R}_{n_2}^2(\mathbf{G_1^D}, \mathbf{G_2^D})$. Second, we design a statistic $\triangle dCor = \frac{|dCor_{n_1}^C - dCor_{n_2}^D|}{dCor_{n_2}^D}$ to measure the difference in distance correlations between cases and controls. This represents how different the two joint distributions $(\boldsymbol{G}_1^C, \boldsymbol{G}_2^C)$ and $(\boldsymbol{G}_1^D, \boldsymbol{G}_2^D)$ are. The larger the $\triangle dCor$, the higher the probability that Gene 1 and Gene 2 interact.

Because we have no information about the distribution of our designed statistic, it is difficult to use a conventional parametric test to do the statistical inference. Therefore, we apply the permutation strategy to estimate the significance of gene–gene interaction. During the permutation test, we rearrange label y to generate a new random case and control label, calculate $\triangle dCor$, construct the empirical distribution, and estimate the p-value. We do the permutation m times and get $\triangle dCor^1, \triangle dCor^2, ..., \triangle dCor^m$. The statistic for the original dataset is $\triangle dCor^0$

Here, the null hypothesis and the alternative hypothesis are defined as follows:

$$H_0 : \triangle dCor^i \text{ has the same distribution}$$

$$H_1 : \triangle dCor^0 \text{ has a distribution different from the other } \triangle\, dCor^i \tag{5}$$

After the permutation, the random samples follow the null hypothesis H_0. According to m statistics from random permutation samples, we can derive the sampling distribution (i.e., empirical distribution) for the statistic $\triangle dCor$ following the null hypothesis H_0.

We count the number of statistics $\triangle dCor^i$ that are equal to or greater than $\triangle dCor^0$.

$$num = \sum_{i=1}^{m} I(\triangle dCor^i \geq \triangle dCor^0)$$

$$I(\triangle dCor^i \geq \triangle dCor^0) = \begin{cases} 1, \triangle dCor^i \geq \triangle dCor^0 \\ 0, \triangle dCor^i < \triangle dCor^0 \end{cases} \tag{6}$$

Then, we estimate the p-value by:

$$p = \frac{num}{m} \tag{7}$$

The framework for GBDcor is described in Algorithm 1.

Algorithm 1: GBDcor

Data: Genotype $\mathbf{G_1}, \mathbf{G_2}$, phenotype $\mathbf{y}$, permutation times m
Result: significance p-value for interaction between $\mathbf{G_1}, \mathbf{G_2}$
Calculate $dCor^C_{n_1}$ and $dCor^D_{n_2}$ for both $(\boldsymbol{G}^C_1, \boldsymbol{G}^C_2)$ and $(\boldsymbol{G}^D_1, \boldsymbol{G}^D_2)$ by Equation (4);
Calculate the difference $\triangle dCor$ between $dCor^C_{n_1}$ and $dCor^D_{n_2}$;
for $i = 1$ *to* m **do**
  Permute the y label, and generate new dataset;
  Repeat Steps 1 and 2;
end
Estimate p-value of $\triangle dCor$

2.2. Simulation Study

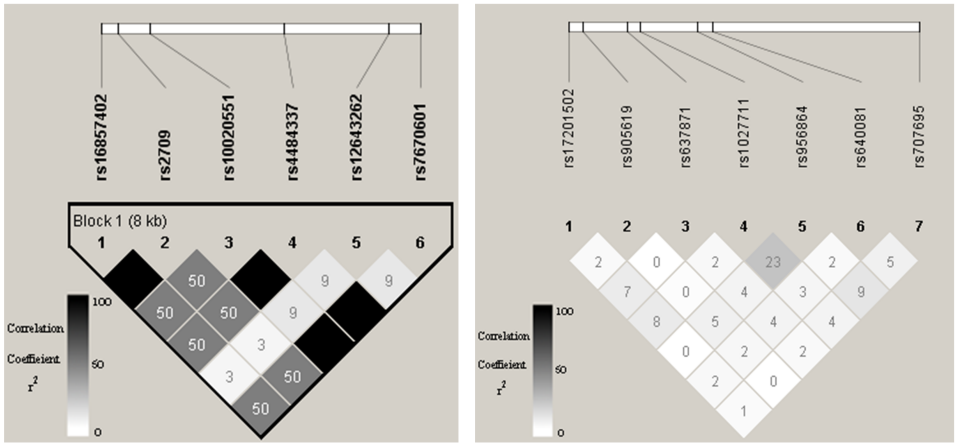
To evaluate the power to detect gene–gene interaction and the ability to control the type-I error rate of GBDcor, we compared GBDcor with three existing techniques: kernel canonical correlation analysis (KCCA) [28,29], the gene-based information gain method (GBIGM) [30], and gene-based gene-gene interaction test (AGGrEGATOr) [31].

2.2.1. Simulation Based on Haplotype Population

Here, we used gs2.0 to generate the simulation data. gs2.0 [38] takes haplotypes as input, then generates dense SNP genotype data for case-control samples that share similar local linkage disequilibrium (LD) patterns as those in human populations. By varying the odds ratio (OR), population prevalence, and sample size, it can generate different disease models. To mimic the real LD structure in a human population, we selected the U.S. Utah residents with ancestry from Northern and Western Europe from the CEPH collection (CEU population) of Hapmap3 (https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html) as template haplotype data. The CEU dataset contains 90 haplotypes. In this study, we chose two gene loci randomly, GNPDA2 on chromosome 4 and FAIM2 on chromosome 12. We imputed chromosome 4 and chromosome 12 using the genipe module, which is a genome-wide imputation pipeline that uses Plink, shapeit, and impute2, with the 1000 Genome Project phase3 data as reference data. After imputation, we got 6 SNPs in GNPDA2 and 7 SNPs in FAIM2 (Table 1 Figure 1).

Table 1. Detailed information about GNPDA2 and FAIM2 used in a study of gene–gene interaction. Shown are the rsID (rs number used by researchers and databases to refer to specific SNPs)and physics position of each SNP on each gene.

Index	SNP Name: Position	
	GNPDA2 (chr4)	**FAIM2 (chr12)**
1	rs16857402:44706453	rs17201502:50285562
2	rs2709:44706913	rs905619:50286055
3	rs10020551:44707815	rs637871:50287592
4	rs4484337:44711547	rs1027711:50288032
5	rs12643262:44714455	rs956864:50290023
6	rs7670601:44715341	rs640081:50290554
7		rs707695:50297670



(a) LD pattern of GNPDA2 (b) LD pattern of FAIM2

Figure 1. Linkage disequilibrium (LD) patterns of GNPDA2 and FAIM2 used in simulation studies. Figures are LD plots produced using Haploview. GNPDA2 has 6 SNPs, and FAIM2 has 7 SNPs. The number in each square is the LD strength that was measured by r^2, where $0 \leq r^2 \leq 1$, 0 means no LD, and 1 means complete LD.

2.2.2. Disease Model

A disease model is a model that expresses the relationship between genes and the disease. Here, we considered two-locus disease models. We take a jointly recessive-recessive model as an example. Suppose population prevalence is p and the genotype odds ratio is $(1+\theta)$ for each locus (Table 2).

Table 2. Odds table of the recessive-recessive model.

SNP1	SNP2		
	BB	**Bb**	**bb**
AA	γ	γ	γ
Aa	γ	γ	γ
aa	γ	γ	$\gamma(1+\theta)$

Let $Pr(D|g_i)$ denote the probability of a sample being affected given the genotype g_i (i.e., the penetrance of g_i), and let $Pr(\bar{D}|g_i)$ denote the probability of a sample not being affected given genotype g_i. Therefore, the odds of a disease can be written as follows:

$$ODD_{g_i} = \frac{Pr(D|g_i)}{Pr(\bar{D}|g_i)} = \frac{Pr(D|g_i)}{1 - Pr(D|g_i)} \tag{8}$$

The penetrance of genotype g_i can be calculated:

$$Pr(g_i) = \frac{ODD_{g_i}}{1 + ODD_{g_i}} \tag{9}$$

Table 3 is the corresponding penetrance table for Table 2.

Table 3. Penetrance table of the recessive-recessive model.

SNP1	SNP2		
	BB	**Bb**	**bb**
AA	$\frac{\gamma}{1+\gamma}$	$\frac{\gamma}{1+\gamma}$	$\frac{\gamma}{1+\gamma}$
Aa	$\frac{\gamma}{1+\gamma}$	$\frac{\gamma}{1+\gamma}$	$\frac{\gamma}{1+\gamma}$
aa	$\frac{\gamma}{1+\gamma}$	$\frac{\gamma}{1+\gamma}$	$\frac{\gamma(1+\theta)}{1+\gamma(1+\theta)}$

Once the population prevalence p and the genotype odds ratio $(1+\theta)$ are fixed in this model, we can calculate the baseline value γ, which represents the odds of disease when the two loci do not carry any disease alleles, by using the following formula and the terms in Table 3.

$$p = Pr(D) = \sum Pr(D|g_i) \times Pr(g_i) \tag{10}$$

We used eight build-in disease models in gs2.0, which included an additive-additive model, recessive-recessive model, threshold model, XOR model, dominant-dominant model, multiplicative model, recessive-dominant model, and a special interaction model. By varying population prevalence, odds ratio, and sample size, we generated different datasets to perform a comparative analysis of AGGrEGATOr, KCCU, and GBIGM.

Type-I error: Type-I error is the probability of rejecting the null hypothesis when the null hypothesis is true. In this paper, we set the significance level at $\alpha = 0.05$. We performed the simulation 100 times with each sample size $n \in \{1k, 2k, 3k, 4k, 5k\}$, by setting the odds ratio at 1.

Power: The power of a statistical test is the probability that it rejects the null hypothesis correctly when the alternative hypothesis is true. In this paper, we ran the simulations 100 times for each parameter combination. The power for each parameter combination is the frequency of rejection of the null hypothesis in the dataset when the alternative hypothesis is true under the significance level of $\alpha = 0.05$. To evaluate the effect of the odds ratio, we varied the odds ratio $OR \in \{1.5, 2, 2.5, 3, 3.5, 4\}$ with population prevalence at $p = 0.01$ and a sample size of $k = 4000$ (2000 cases and 2000 controls). To evaluate the effect of sample size, we choose $n \in \{1k, 2k, 3k, 4k, 5k\}$ with an odds ratio of $OR = 2$ and population prevalence of $p = 0.01$.

For GBDcor, AGGrEGATOr, KCCU, and GBIGM, if the number of datasets with a p-value less than α was m_1, the power was calculated by:

$$power = \frac{m_1}{100} \tag{11}$$

For GBDcor, AGGrEGATOr, and GBIGM, we used a nonparametric method with no parameter specified. For KCCU, we set the ratio for a trimmed jackknife at 0.05 ($\omega = 0.05$).

2.3. Application with Rheumatoid Arthritis Data

To assess the capacity of GBDcor to deal with real gene–gene interaction of a case-control dataset, we investigated the susceptibility of a set of pair of genes in rheumatoid arthritis (RA), which is a chronic, autoimmune joint disease where persistent inflammation affects bone remodeling and results in progressive bone destruction. We used the WTCCC (2007) dataset, which was genotyped in a British population using the Affymetrix GeneChip 500k.

To verify our method, we constructed our dataset in the following ways:

(1) We wanted to verify some gene–gene interaction in the RA pathway hsa05323 in the KEGG pathway dataset. Genotyping coordinates are given in NCBI Build36/UCSC hg18 (National Center for Biotechnology Information, Bethesda, MD, USA). There is a total of 90 genes in this pathway. Because MHCII and V-ATPase are two protein combinations with many interactions within themselves, we only chose a representative gene from each of them and excluded other

genes. After that, 48 genes were left. Each unique gene pair was evaluated, which resulted in a total of $\binom{48}{2} = 1128$ pairs for those genes.

(2) We obtained gene information from the NCBI Build36 annotation file. For each gene, 10 kb of buffer region were added both upstream and downstream of the defined gene location. All SNPs between the regions were considered.

(3) Based on the quality control provided by GWAS, we removed samples where the reported sex did not match the heterozygosity rates observed on chromosome X. We also excluded an SNP if its minor allele frequency (MAF) was <0.05, if its missing rate was >10% of the samples, or if its frequencies in the control were not in Hardy–Weinberg equilibrium ($p < 0.0001$). After filtering, there were 385 SNPs left in 4966 samples, which consisted of 1973 cases and 2993 controls.

3. Results

3.1. Simulation Study

3.1.1. Type-I Error

After we set the significance level at $\alpha = 0.05$, changing the sample size gradually resulted in type-I errors for all the methods that were close to the significance level for most sample size settings (Table 4), except for GBIGM at $n = 1k$. The type-I error was controlled by these methods with different sample sizes with no effects.

Table 4. Type-I error of the four methods in different sample sizes. AGGrEGATOr, a gene-based gene gene interaction; GBDcor, gene-based gene-gene interaction via distance correlation coefficient GBIGM, gene-based information gain method; KCCU, kernelized CCU.

Method	Sample Size				
	1*k*	**2*k***	**3*k***	**4*k***	**5*k***
AGGrEGATOr	0.05	0.06	0.07	0.04	0.02
GBDcor	0.05	0.03	0.04	0.04	0.06
GBIGM	0.13	0.06	0.07	0.07	0.07
KCCU	0.02	0.02	0.01	0.05	0.07

3.1.2. Power

The effect of the odds ratio: We assessed the performance in detecting gene–gene interaction under eight disease models. The curves were constructed while varying the odds ratio ($OR \in \{1.5, 2, 2.5, 3, 3.5, 4\}$) with population prevalence set at 0.01 and sample size set at 4k (Figure 2). Notice that a larger power indicated better performance. For this experiment, we chose one pair of SNPs belonging to different genes randomly to be causal to generate the simulated dataset. We considered the two genes that contain the SNPs to be interacting. The performance of all methods improved when OR became larger, and the power tended to be one for some methods at $OR = 4$. Among them, GBDcor was the best, except for the additive-additive model (AA model). GBIGM showed the best performance under this model; however, it has been declared that GBIGM had a fatal inflation problem under this disease model. We also noticed that for the recessive-recessive model (RR model), when the OR value changed gradually from 1.5–4, the power was consistently $\leq 20\%$. AGGrEGATOr reached 40%, and GBDcor was approximately 60%. According to the penetrance table for the recessive-recessive model (Table 3), when we set population prevalence $p = 0.01$, the baseline γ was a very small number. Therefore, among nine genotypes, eight of them tended to be zero. The only genotype (aabb) that was causal consisted of two minor alleles. Usually, the minor allele frequency of SNP was 0.2–0.4, which caused the genotype (aabb) to emerge only barely in the simulation dataset. Therefore, it was difficult to see any difference between cases and controls. That is, these methods showed poor performance under this model.

The effect of sample size: Next, we explored the impact of sample size. We set sample size $n \in \{1k, 2k, 3k, 4k, 5k\}$ with $OR = 2$ and $p = 0.01$ (Figure 3). With increasing sample size, the power of all the methods increased monotonically under all disease models, except the RR model. Other than GBIGM, GBDcor performed much better than KCCU or AGGrEGATOr under the AA model. The power of GBDcor reached 60%, but the other two methods were $\leq$ 30%. For all methods, larger sample size led to better performance.

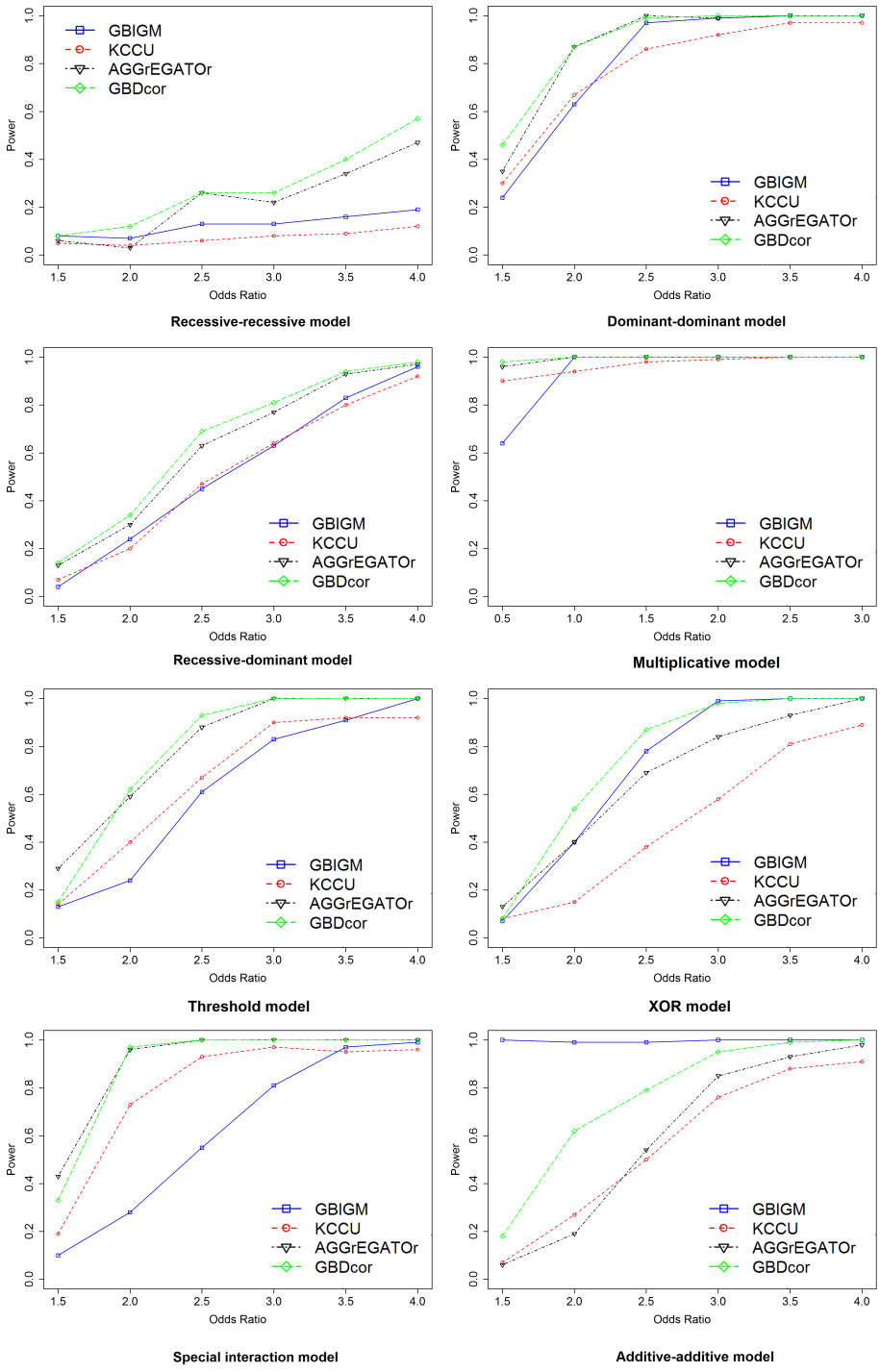


Figure 2. Empirical, simulation-based statistical power of GBIGM, KCCU, AGGrEGATOr, and GBDcor under eight disease models, after varying the $OR \in \{1.5, 2, 2.5, 3, 3.5, 4\}$.

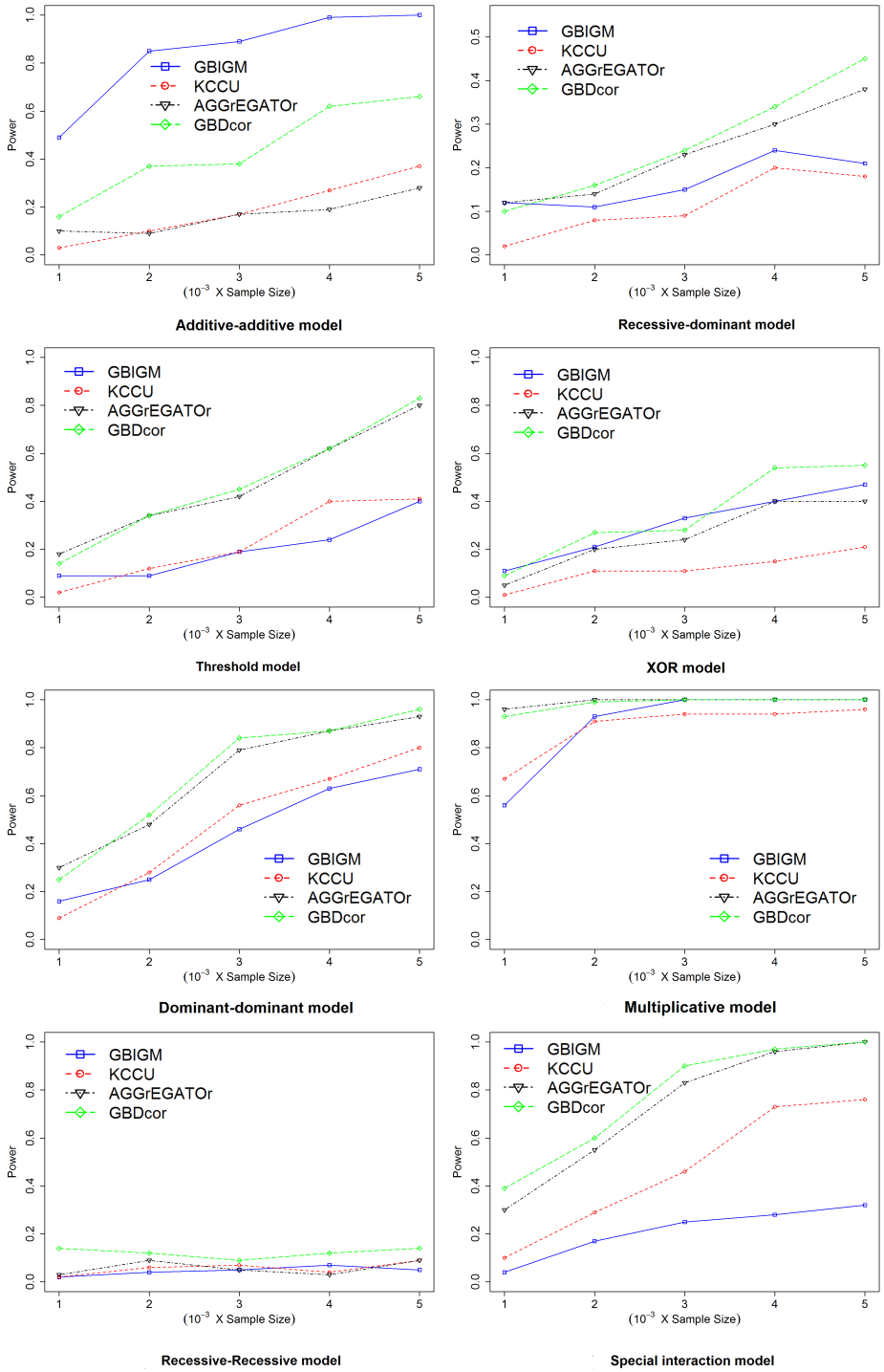


Figure 3. Empirical, simulation-based statistical power of GBIGM, KCCU, AGGrEGATOr, and GBDcor under eight disease models, after varying the $n \in \{1k, 2k, 3k, 4k, 5k\}$.

GBDcor performed better than alternative methods for almost the entire range of settings that we used. The benefits of using distance correlation to learn the dependence relationship of two genes in cases and controls were pronounced in the gene–gene interaction detection scenario. For example, we were able to design a statistic to represent the degree of difference of the two distance correlation coefficients and to apply a permutation to find the empirical distribution of our designed statistic.

3.2. Application Using Rheumatoid Arthritis Data

Rheumatoid arthritis (RA) is an autoimmune synovitis characterized by the formation of panus and destruction of cartilage and bone in synovial joints. TNF-α, IL-6, IL-17, MMPs, and RANK are some of the main players in the development of RA [39]. For the RA study of the hsa05323 pathway, we obtained 1128 pairs of genes to evaluate. For our method, we did permutation $m = 1000$ times. Using a significance level of $\alpha = 0.01$, KCCU and GBIGM obtained 159 and 134 significant gene–gene interaction (GGIs), respectively. Thirty and 65 had a p-value equal to zero, respectively. AGGrEGATOr had 17 significant GGIs, and GBDcor had 18 significant GGIs.

Because GBIGM and KCCU had too many pairs, we were unable to analyze all of them. We selected the top 10 in GBDcor and in AGGrEGATOr for analysis. Then, we listed the p-value for each of the 20 pairs of genes for each of the methods (Table 5). We found seven of 10 results for GBDcor in the literature that supported our results, and we found two of 10 for AGGrEGATOr that did so. The column 'Ref' in Table 5 gives the references for the literature evidence that show direct interaction between two genes. We also observed that there were more intersections among GBDcor, KCCU, and GBIGM than among AGGrEGATOr, KCCU, and GBIGM.

Table 5. The p-values of the gene pairs detected to interact from different methods. The p-values with bold font mean they are significant

Gene1	Gene2	Ref	*p*-Value			
			GBDcor	AGGrEGATOr	KCCU	GBIGM
AP-1	M-CSF	ref [40]	**0**	0.0679	**0.001**	**0**
CXCL12	FLT-1		**0**	0.59	0.152	**0**
GM-CSF	VEGF	ref [41]	**0.001**	0.284	**0.005**	0.545
CTSK	VEGF		**0.002**	0.873	**0.028**	0.47
CTLA4	TLR2		**0.002**	0.152	0.057	**0.008**
CXCL1	RANK	ref [42–44]	**0.002**	**0.024**	0.147	0.697
IL15	MMP-3	ref [45]	**0.002**	0.066	0.167	0.088
GM-CSF	AP-1	ref [46,47]	**0.002**	0.394	**0.001**	**0.027**
CD86	APRIL	ref [48]	**0.003**	0.637	**0.03**	0.655
TGFβ	VEGF	ref [40]	**0.005**	1	**0.029**	0.632
CD80	APRIL	ref [48]	0.865	**0.0006**	0.941	0.334
CTSK	BLyS		0.298	**0.0008**	0.356	0.056
AP-1	IL-6		0.24	**0.0018**	0.098	0.287
CD80	CTSL		0.094	**0.0019**	0.519	0.252
CXCL6	FLT-1		0.441	**0.0023**	**0.004**	0.52
CTLA4	AP-1		0.075	**0.0023**	**0.042**	0.102
FLT1	LFA-1		0.645	**0.0031**	0.063	**0.028**
CCL3	TRAP		0.746	**0.0032**	0.682	**0**
IL-18	TGFβ		0.841	**0.0036**	0.149	0.22
IL-1	SDF-1	ref [49]	0.618	**0.004**	0.116	0.636

4. Conclusions

Case-control datasets are common and important in research in medicine and evolutionary biology. In this paper, we developed a gene-based, gene–gene interaction detection algorithm called GBDcor that was based on distance correlation coefficients and a permutation strategy for GWAS on case-control datasets. The method benefits from the ability of distance correlation coefficients, which can detect nonlinear models, and the robustness of our hypothesis testing scheme, which is based on permutation and is non-parametric.

As a consequence, GBDcor was able to detect interpretable gene–gene interaction more accurately and effectively compared to other methods. We demonstrated such effectiveness through a semi-empirical simulation study and retrospective analysis of rheumatoid arthritis. Under a large range of settings, GBDcor outperformed previous methods in the power of detecting gene–gene

interaction. The method was also stable to sample size based on a test of false positive rates. The distance correlation had no limitation on the dimension of two random vectors. Therefore, it is possible to generalize the method for pairwise, marker-based detection of gene pairs that were identified as interactive. Investigating the mechanism of gene-level interaction at the marker-level might be a direction for further research. In summary, GBDcor is a useful addition to the current toolbox of statistical models for unraveling gene–gene interaction in case-control studies.

Author Contributions: Conceptualization, Y.G.; formal analysis, Y.G.; funding acquisition, Maozu Guo and X.L.; methodology, Y.G. and C.W.; project administration, M.G.; software, Y.G.; writing, original draft, Y.G.; writing, review and editing, Y.G., C.W., M.G., X.L., and A.K.

Funding: This work was supported by the National Natural Science Foundation of China (Grant Nos. 61571163, 61532014, and 61671189) and the National Key Research and Development Plan of China (Grant No. 2016YFC0901902).

Acknowledgments: We would like to thank Yingyuan Guo, The Second Hospital of Jilin University, for help with analyzing the real data results. We also thank the members of the Natural Computing group for thoughtful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J.P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367. [CrossRef]
2. MacArthur, J.; Bowler, E.; Cerezo, M.; Gil, L.; Hall, P.; Hastings, E.; Junkins, H.; McMahon, A.; Milano, A.; Morales, J.; et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **2017**, *45*, D896–D901. [CrossRef] [PubMed]
3. Fadista, J.; Lund, M.; Skotte, L.; Geller, F.; Nandakumar, P.; Chatterjee, S.; Matsson, H.; Granström, A.L.; Wester, T.; Salo, P.; et al. Genome-wide association study of Hirschsprung disease detects a novel low-frequency variant at the RET locus. *Eur. J. Hum. Genet.* **2018**, *26*, 561. [CrossRef] [PubMed]
4. Xu, Y.; Wang, Y.; Luo, J.; Zhao, W.; Zhou, X. Deep learning of the splicing (epi) genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision. *Nucleic Acids Res.* **2017**, *45*, 12100–12112. [CrossRef] [PubMed]
5. Xu, Y.; Zhao, W.; Olson, S.D.; Prabhakara, K.S.; Zhou, X. Alternative splicing links histone modifications to stem cell fate decision. *Genome Biol.* **2018**, *19*, 133. [CrossRef]
6. Cordell, H.J. Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.* **2009**, *10*, 392. [CrossRef]
7. Moore, J.H.; Asselbergs, F.W.; Williams, S.M. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* **2010**, *26*, 445–455. [CrossRef] [PubMed]
8. Manolio, T.A.; Collins, F.S.; Cox, N.J.; Goldstein, D.B.; Hindorff, L.A.; Hunter, D.J.; McCarthy, M.I.; Ramos, E.M.; Cardon, L.R.; Chakravarti, A.; et al. Finding the missing heritability of complex diseases. *Nature* **2009**, *461*, 747. [CrossRef]
9. Moore, J.H.; Williams, S.M. Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.* **2009**, *85*, 309–320. [CrossRef]
10. Zuk, O.; Hechter, E.; Sunyaev, S.R.; Lander, E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 1193–1198. [CrossRef]
11. Marchini, J.; Donnelly, P.; Cardon, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **2005**, *37*, 413. [CrossRef] [PubMed]
12. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; De Bakker, P.I.; Daly, M.J.; et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef]
13. Lin, H.; Mueller-Nurasyid, M.; Smith, A.V.; Arking, D.E.; Barnard, J.; Bartz, T.M.; Lunetta, K.L.; Lohman, K.; Kleber, M.E.; Lubitz, S.A.; et al. Gene-gene Interaction Analyses for Atrial Fibrillation. *Sci. Rep.* **2016**, *6*, 35371. [CrossRef] [PubMed]

14. Emily, M. IndOR: A new statistical procedure to test for SNP–SNP epistasis in genome-wide association studies. *Stat. Med.* **2012**, *31*, 2359–2373. [CrossRef] [PubMed]
15. Zhao, J.; Jin, L.; Xiong, M. Test for interaction between two unlinked loci. *Am. J. Hum. Genet.* **2006**, *79*, 831–845. [CrossRef]
16. Wu, X.; Dong, H.; Luo, L.; Zhu, Y.; Peng, G.; Reveille, J.D.; Xiong, M. A novel statistic for genome-wide interaction analysis. *PLoS Genet.* **2010**, *6*, e1001131. [CrossRef]
17. Ueki, M.; Cordell, H.J. Improved statistics for genome-wide interaction analysis. *PLoS Genet.* **2012**, *8*, e1002625. [CrossRef]
18. Dong, C.; Chu, X.; Wang, Y.; Wang, Y.; Jin, L.; Shi, T.; Huang, W.; Li, Y. Exploration of gene–gene interaction effects using entropy-based methods. *Eur. J. Hum. Genet. EJHG* **2008**, *16*, 229. [CrossRef]
19. Kang, G.; Yue, W.; Zhang, J.; Cui, Y.; Zuo, Y.; Zhang, D. An entropy-based approach for testing genetic epistasis underlying complex diseases. *J. Theor. Biol.* **2008**, *250*, 362–374. [CrossRef]
20. Ritchie, M.D.; Hahn, L.W.; Moore, J.H. Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet. Epidemiol.* **2003**, *24*, 150–157. [CrossRef]
21. Moore, J.H.; White, B.C. Tuning ReliefF for genome-wide genetic analysis. In *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 166–175.
22. Zhang, X.; Huang, S.; Zou, F.; Wang, W. TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics* **2010**, *26*, i217–i227. [CrossRef] [PubMed]
23. Wan, X.; Yang, C.; Yang, Q.; Xue, H.; Fan, X.; Tang, N.L.; Yu, W. BOOST: A fast approach to detecting gene–gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.* **2010**, *87*, 325–340. [CrossRef] [PubMed]
24. Li, J.; Malley, J.D.; Andrew, A.S.; Karagas, M.R.; Moore, J.H. Detecting gene–gene interactions using a permutation-based random forest method. *BioData Min.* **2016**, *9*, 14. [CrossRef] [PubMed]
25. Li, M.X.; Gui, H.S.; Kwan, J.S.; Sham, P.C. GATES: A rapid and powerful gene-based association test using extended Simes procedure. *Am. J. Hum. Genet.* **2011**, *88*, 283–293. [CrossRef] [PubMed]
26. Liu, J.Z.; Mcrae, A.F.; Nyholt, D.R.; Medland, S.E.; Wray, N.R.; Brown, K.M.; Hayward, N.K.; Montgomery, G.W.; Visscher, P.M.; Martin, N.G.; et al. A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **2010**, *87*, 139–145. [CrossRef] [PubMed]
27. Peng, Q.; Zhao, J.; Xue, F. A gene-based method for detecting gene–gene co-association in a case–control association study. *Eur. J. Hum. Genet.* **2010**, *18*, 582. [CrossRef] [PubMed]
28. Yuan, Z.; Gao, Q.; He, Y.; Zhang, X.; Li, F.; Zhao, J.; Xue, F. Detection for gene–gene co-association via kernel canonical correlation analysis. *BMC Genet.* **2012**, *13*, 83. [CrossRef] [PubMed]
29. Larson, N.B.; Jenkins, G.D.; Larson, M.C.; Vierkant, R.A.; Sellers, T.A.; Phelan, C.M.; Schildkraut, J.M.; Sutphen, R.; Pharoah, P.P.; Gayther, S.A.; et al. Kernel canonical correlation analysis for assessing gene–gene interactions and application to ovarian cancer. *Eur. J. Hum. Genet.* **2014**, *22*, 126. [CrossRef] [PubMed]
30. Li, J.; Huang, D.; Guo, M.; Liu, X.; Wang, C.; Teng, Z.; Zhang, R.; Jiang, Y.; Lv, H.; Wang, L. A gene-based information gain method for detecting gene–gene interactions in case-control studies. *Eur. J. Hum. Genet. EJHG* **2015**, *23*, 1566–1572. [CrossRef] [PubMed]
31. Emily, M. AGGrEGATOr: A Gene-based GEne-Gene interActTiOn test for case-control association studies. *Stat. Appl. Genet. Mol. Biol.* **2016**, *15*, 151–171. [CrossRef] [PubMed]
32. Ma, L.; Clark, A.G.; Keinan, A. Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genet.* **2013**, *9*, e1003321. [CrossRef] [PubMed]
33. Székely, G.J.; Rizzo, M.L.; Bakirov, N.K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **2007**, *35*, 2769–2794. [CrossRef]
34. Székely, G.J.; Rizzo, M.L. The distance correlation *t*-test of independence in high dimension. *J. Multivar. Anal.* **2013**, *117*, 193–213. [CrossRef]
35. Szekely, G.J.; Rizzo, M.L. Partial distance correlation with methods for dissimilarities. *Ann. Stat.* **2014**, *42*, 2382–2412. [CrossRef]
36. Zhang, Q. A powerful nonparametric method for detecting differentially co-expressed genes: Distance correlation screening and edge-count test. *BMC Syst. Biol.* **2018**, *12*, 58. [CrossRef] [PubMed]

37. Fang, J.; Xu, C.; Zille, P.; Lin, D.; Deng, H.W.; Calhoun, V.D.; Wang, Y.P. Fast and Accurate Detection of Complex Imaging Genetics Associations Based on Greedy Projected Distance Correlation. *IEEE Trans. Med. Imaging* **2018**, *37*, 860–870. [CrossRef] [PubMed]
38. Li, J.; Chen, Y. Generating samples for association studies based on HapMap data. *BMC Bioinform.* **2008**, *9*, 44. [CrossRef] [PubMed]
39. Cope, A.P. T cells in rheumatoid arthritis. *Arthritis Res. Ther.* **2008**, *10*, S1, doi:10.1186/ar2412. [CrossRef]
40. Shiozawa, S.; Tsumiyama, K. Pathogenesis of rheumatoid arthritis and c-Fos/AP-1. *Cell Cycle* **2009**, *8*, 1539–1543. [CrossRef]
41. Zhao, J.; Chen, L.; Shu, B.; Tang, J.; Zhang, L.; Xie, J.; Qi, S.; Xu, Y. Granulocyte/macrophage colony-stimulating factor influences angiogenesis by regulating the coordinated expression of VEGF and the Ang/Tie system. *PLoS ONE* **2014**, *9*, e92691. [CrossRef]
42. Kirkham, B.W.; Kavanaugh, A.; Reich, K. Interleukin-17A: A unique pathway in immune-mediated diseases: Psoriasis, psoriatic arthritis and rheumatoid arthritis. *Immunology* **2014**, *141*, 133–142. [CrossRef]
43. Mori, T.; Miyamoto, T.; Yoshida, H.; Asakawa, M.; Kawasumi, M.; Kobayashi, T.; Morioka, H.; Chiba, K.; Toyama, Y.; Yoshimura, A. IL-1β and TNFα-initiated IL-6–STAT3 pathway is critical in mediating inflammatory cytokines and RANKL expression in inflammatory arthritis. *Int. Immunol.* **2011**, *23*, 701–712. [CrossRef]
44. Ahn, J.K.; Huang, B.; Bae, E.K.; Park, E.J.; Hwang, J.W.; Lee, J.; Koh, E.M.; Cha, H.S. The role of α-defensin-1 and related signal transduction mechanisms in the production of IL-6, IL-8 and MMPs in rheumatoid fibroblast-like synoviocytes. *Rheumatology* **2013**, *52*, 1368–1376. [CrossRef] [PubMed]
45. Chan, A.; Filer, A.; Parsonage, G.; Kollnberger, S.; Gundle, R.; Buckley, C.D.; Bowness, P. Mediation of the proinflammatory cytokine response in rheumatoid arthritis and spondylarthritis by interactions between fibroblast-like synoviocytes and natural killer cells. *Arthritis Rheum. Off. J. Am. Coll. Rheumatol.* **2008**, *58*, 707–717. [CrossRef] [PubMed]
46. Shiomi, A.; Usui, T. Pivotal roles of GM-CSF in autoimmunity and inflammation. *Mediat. Inflamm.* **2015**, *2015*, 568543. [CrossRef] [PubMed]
47. Johnson, B.V.; Bert, A.G.; Ryan, G.R.; Condina, A.; Cockerill, P.N. Granulocyte-macrophage colony-stimulating factor enhancer activation requires cooperation between NFAT and AP-1 elements and is associated with extensive nucleosome reorganization. *Mol. Cell. Biol.* **2004**, *24*, 7914–7930. [CrossRef] [PubMed]
48. Finnegan, A.; Ashaye, S.; Hamel, K.M. B effector cells in rheumatoid arthritis and experimental arthritis. *Autoimmunity* **2012**, *45*, 353–363. [CrossRef] [PubMed]
49. Zheng, X.; Zhao, F.C.; Pang, Y.; Li, D.Y.; Yao, S.C.; Sun, S.S.; Guo, K.J. Downregulation of miR-221-3p contributes to IL-1β-induced cartilage degradation by directly targeting the SDF1/CXCR4 signaling pathway. *J. Mol. Med.* **2017**, *95*, 615–627. [CrossRef] [PubMed]