Zhang et al. *BMC Medical Genomics* 2019, **12**(Suppl 7):133
https://doi.org/10.1186/s12920-019-0583-7

BMC Medical Genomics

METHODOLOGY Open Access

Region-based interaction detection in genome-wide case-control studies

Sen Zhang[1], Wei Jiang[2], Ronald CW Ma[3] and Weichuan Yu[2*]

From 14th International Symposium on Bioinformatics Research and Applications (ISBRA'18)
Beijing, China. 8-11 June 2018

Abstract

Background: In genome-wide association study (GWAS), conventional interaction detection methods such as BOOST are mostly based on SNP-SNP interactions. Although single nucleotides are the building blocks of human genome, single nucleotide polymorphisms (SNPs) are not necessarily the smallest functional unit for complex phenotypes. Region-based strategies have been proved to be successful in studies aiming at marginal effects.

Methods: We propose a novel region-region interaction detection method named RRIntCC (region-region interaction detection for case-control studies). RRIntCC uses the correlations between individual SNP-SNP interactions based on linkage disequilibrium (LD) contrast test.

Results: Simulation experiments showed that our method can achieve a higher power than conventional SNP-based methods with similar type-I-error rates. When applied to two real datasets, RRIntCC was able to find several significant regions, while BOOST failed to identify any significant results. The source code and the sample data of RRIntCC are available at http://bioinformatics.ust.hk/RRIntCC.html.

Conclusion: In this paper, a new region-based interaction detection method with better performance than SNP-based interaction detection methods has been proposed.

Keywords: GWAS, Statistical interaction detection, Region-based method, LD contrast test

Background

Genome-wide association study (GWAS) has served as an important tool to investigate the relationship between genomic variants and human traits [1]. The genetic variants investigated in GWAS are mainly single nucleotide polymorphisms (SNPs). SNPs are single nucleotide variants whose genotypes are not fixed in the population and exhibit diversities among different individuals. Most GWAS analysis protocols follow the single-locus test procedures aimed at detecting the marginal effects of SNPs [2, 3]. However, it's well recognized that genetic variants work synergistically through certain pathogenic pathways [4]. The interactions among SNPs are not guaranteed to be discovered by marginal effect detection, especially for SNPs with weak marginal effects but strong interaction effects [5]. Many methods have been developed to address this problem [4, 6], including PLINK [7], BOOST [5], MDR [8], ReliefF [9], BEAM [10], and LD contrast test [11].

An important problem in SNP-SNP interaction detection is the stringent threshold when considering multiple testing correction. For marginal effect detection, a SNP can only be considered as significant when its corresponding p-value is at the order of 10^{-8} (assuming we use Bonferroni correction). In SNP-SNP interaction detection, the threshold has to go down further to the order of 10^{-14}. As a result, interactions with weak or moderate effect sizes might remain undiscovered.

In this paper, we proposed a region-based interaction detection method to address this problem. Region-based methods have been successful in marginal effect detection [12, 13]. The basic idea is to group the effects of nearby SNPs together and test their aggregation rather than investigating the elements separately. The benefit

*Correspondence: eeyu@ust.hk
[2]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Kowloon, Hong Kong, China
Full list of author information is available at the end of the article

is two-folds: Firstly, the size and the number of regions are controllable. We can achieve the balance between the resolution of the results and the statistical significance threshold after Bonferroni correction. Secondly, the effect size might be enhanced by taking the whole region into account. SNPs are the basic genomic units. Neverthless, they are not necessarily the functional units of diseases. Different SNP mutations in a gene can all lead to changes of protein functions. Therefore, grouping different SNPs together provides a possible alternative.

To group different SNP-SNP pairs together, the key is to quantitatively measure and account for the relationships between different SNP-SNP pairs. To the best of our knowledge, no existing method is available to test region-region interactions for case-control studies, where we only have two groups of people: healthy people (controls) and people with the investigated disease (cases). Although Ma et al. [14] proposed a region-based interaction detection method to analyze continuous traits based on the linear regression model, it is not easy to extend their method to the case-control setting due to the difficulty of deriving the covariances of test statistics under the logistic regression model that is commonly used in case-control studies. In this paper, we use the LD contrast test method instead of the logistic regression in interaction detection. We derive the correlation coeffcients of the correpsonding SNP-SNP interaction test statistics. Then we further extend region-based methods to the case-control setting by accounting for the covariances between SNP-based test statistics. We name this method RRIntCC (region-region interaction detection for case-control studies). Experiment results illustrate that RRIntCC achieves a higher power than conventional SNP-SNP interaction detection methods at the same type-I-error rate.

Methods

Here we propose a novel region-based interaction detection method for genome-wide case-control studies that utilizes SNP-based interaction test statistics and their covariances. LD contrast test is adopted to measure SNP-based interaction effects. We derive the covariance of LD contrast test statistics, which enables a robust aggregation of SNP-SNP interactions within a region pair. The determination of regions comes from gene definitions or BOOST results.

Genomic data formats

There are two alleles for almost every base pair (bp) position in the human genome, one from the maternal chromosome and the other from the paternal chromosome. A combination of the two alleles is denoted as a genotype of this bp position. SNPs are defined as the base pairs that could exhibit different genotype values in different individuals. Normally a SNP only has two possible allele values in the population, one major allele with a higher probability (denoted as B), and one minor allele (denoted as b). Correspondingly, there exist three genoytpes for a typical SNP, i.e., BB, Bb and bb, where Bb is called a heterogeneous genotype and the rest two are called homogeneous genotypes. GWAS uses microarrays to generate SNP genotype data. In SNP data analysis, we use 0/1/2, 0/1/1, and 0/0/1 for BB/Bb/bb as the encoding scheme for additive, dominant, and recessive genetic models, respectively. A more flexible strategy is to estimate the effects of three genotypes independently, at the price of an increased degree of freedom. Allele data could also be used for analysis, with 0/1 as the numerical values of major/minor alleles. However, statistical inference needs to be performed in advance to retrieve allele information from original genotype data, which is called haplotype phasing in the GWAS community. In this paper, we focus on the analysis of genotype data.

LD contrast test for SNP interaction detection

Current interaction detection methods are mainly based on the deviation from additive effect by assuming a linear or logistic regression model. Nevertheless, this approach is not necessarily the most powerful method due to the uncertainty of underpinning biochemical mechanisms. Linkage disequilibrium (LD) contrast test provides another valuable perspective to investigate this problem. Empirical studies have shown that LD contrast test can achieve higher power than logistic regression under certain disease models for case-control studies [6]. In this paper, LD contrast test is adopted to generate SNP-based interaction test statistics because of its clear statistical meaning and mathematical simplicity.

LD represents the statistical association between two genetic loci with allele values, defined as the deviation from the independence of two SNPs (A and B)

$$LD = p(A,B) - p(A)p(B). \tag{1}$$

To avoid the ambiguity caused by haplotype phasing, composite LD (CLD) which only requires genotype data is commonly used to approximate LD. CLD is defined as [15]:

$$\begin{aligned} CLD &= p_{AB} + p'_{AB} - 2p(A)p(B) \\ with \quad &\begin{cases} p_{AB} = P_{AB}^{AB} + \frac{1}{2}\left(P_{Ab}^{AB} + P_{aB}^{AB} + P_{ab}^{AB}\right) \\ p'_{AB} = P_{AB}^{AB} + \frac{1}{2}\left(P_{Ab}^{AB} + P_{aB}^{AB} + P_{Ab}^{aB}\right) \end{cases}, \end{aligned} \tag{2}$$

where the subscript and the superscript represent two gametes that are passed to offsprings and P denotes the probability of the specific gamete combination. CLD could be regarded as a simplified version of phasing to facilitate the analysis based on genotype data. The statistical properties of CLD have been well studied [16, 17]. One important fact is that CLD corresponds to the sample correlation coefficient $\hat{r}$ of genotype values under the additive model,

$$\hat{r}_{genotype} = \frac{CLD}{\sqrt{p(1-p)+D_A}\sqrt{q(1-q)+D_B}} \approx \frac{CLD}{\sqrt{p(1-p)}\sqrt{q(1-q)}}, \tag{3}$$

where $p = p(A)$, $q = p(B)$, D_A and D_B represent Hardy-Weinberg disequilibriums, i.e. $D_A = p_{AA} - p^2(A)$, $D_B = p_{BB} - p^2(B)$. D_A and D_B are nearly 0 in GWAS datasets after quality control.

A similar result holds for the original LD and allele values,

$$\hat{r}_{allele} = \frac{LD}{\sqrt{p(1-p)}\sqrt{q(1-q)}}. \tag{4}$$

Therefore, CLD could also be viewed as an approximation of LD by using the correlation coefficient of 0/1/2 genotype data under the addtive model to replace that of 0/1 allele values, at the price of implicitly conducting phasing with equal probabilities for two-allele combinations.

Suppose two SNPs work synergistically to contribute to the same pathways, they are less likely to be separated during recombination and will be inherited together to offsprings in the case group. As a result, the SNP-SNP pattern should be different between patients and healthy people. Therefore, checking the difference of LD patterns between cases and controls provides an alternative way to detect interaction. LD contrast test was proposed to statistically test this difference [11]. The test statistic based on CLD has the following form:

$$\chi^2 = \frac{\left(\hat{CLD}_{AB}^{case} - \hat{CLD}_{AB}^{control}\right)^2}{Var\left(\hat{CLD}_{AB}^{case}\right) + Var\left(\hat{CLD}_{AB}^{control}\right)}, \tag{5}$$

which follows a 1-df χ^2 distribution under the null hypothesis that there is no LD difference between cases and controls.

Covariance between SNP interactions

The key issue in the aggregation of individual SNP-SNP interaction effects is the correction of inflated effect sizes caused by the correlations among individual test statistics. The fact that LD is actually the sample covariance of two SNPs is leveraged to derive the correlation coefficients of LD contrast test statistics.

Suppose two SNP pairs (X, Y) and (U, V) have interactions with contrast LDs

$$\begin{cases} \Delta LD_{XY} = \hat{cov}(X, Y|case) - \hat{cov}(X, Y|control) \\ \Delta LD_{UV} = \hat{cov}(U, V|case) - \hat{cov}(U, V|control) \end{cases}. \tag{6}$$

The corresponding LD contrast test statistics read:

$$T_{XY} = \frac{\Delta\hat{LD}_{XY}}{\sqrt{Var(\Delta\hat{LD}_{XY})}} \text{ and } T_{UV} = \frac{\Delta\hat{LD}_{UV}}{\sqrt{Var(\Delta\hat{LD}_{UV})}}. \tag{7}$$

The covariance of the two test statistics reads:

$$cov(T_{XY}, T_{UV}) \approx \frac{cov(\Delta\hat{LD}_{XY}, \Delta\hat{LD}_{UV})}{\sqrt{cov(\Delta\hat{LD}_{XY}, \Delta\hat{LD}_{XY})cov(\Delta\hat{LD}_{UV}, \Delta\hat{LD}_{UV})}}. \tag{8}$$

In GWAS, it's commonly assumed that population samples are independent. Under this assumption, we can derive the following theorems.

Theorem 1. *The covariance of contrast LDs can be decomposed into components from cases and controls separately,*

$$\begin{aligned} cov(\Delta LD_{XY}, \Delta LD_{UV}) =& cov[\,\hat{cov}(X, Y|case), \hat{cov}(U, V|case)] \\ &+ cov[\,\hat{cov}(X, Y|control), \hat{cov}(U, V|control)]\,. \end{aligned} \tag{9}$$

Proof 1. *ΔLD is the difference of the two sample covariances in cases and controls. By the linear property of covariance, $cov(\Delta LD_{XY}, \Delta LD_{UV})$ can be decomposed into four covariances of two sample covariances. Because individuals are assumed to be independent, the two terms with one sample covariance from cases and the other from controls are 0. Therefore, Theorem 1 holds.* □

Theorem 2. *The covariance of sample covariances reads*

$$cov\left[\hat{cov}(X, Y), \hat{cov}(U, V)\right] = \frac{1}{n}\left(\delta_4 - \delta_2 + \frac{\sigma_2 + \tau_2}{n-1}\right), \tag{10}$$

where

$$\begin{aligned} \delta_4 &= E\left[(X - EX)(Y - EY)(U - EU)(V - EV)\right], \\ \delta_2 &= cov(X, Y)cov(U, V), \\ \sigma_2 &= cov(X, U)cov(Y, V), \\ \tau_2 &= cov(X, V)cov(Y, U). \end{aligned}$$

Proof 2. *The covariance of sample covariances can be rewritten as*

$$\begin{aligned} &cov\left[\hat{cov}(X, Y), \hat{cov}(U, V)\right] \\ &= cov\left[\frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)(Y_i - Y_j), \frac{1}{2n(n-1)}\right. \\ &\left.\sum_{i=1}^{n}\sum_{j=1}^{n}(U_i - U_j)(V_i - V_j)\right] \\ &= \frac{1}{4n^2(n-1)^2}\sum_{j=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \\ &cov\left[(X_i - X_j)(Y_i - Y_j),\ (U_k - U_l)(V_k - V_l)\right]. \end{aligned} \tag{11}$$

We consider the following four conditions. (1) $i = j$ or $k = l$. (2) $i \neq j, i \neq k, i \neq l, j \neq k, j \neq l$ and $k \neq l$. (3) $i \neq j$ and $\{i = k, j = l$ or $i = l, j = k\}$. (4) $i \neq j, k \neq l$, and $\{i = k$ or

$i = l$ or $j = k$ or $j = l$}. The basic covariance unit in (11) can be rewriten as

$$cov\left\{\left[(X_i - EX) - (X_j - EX)\right]\left[(Y_i - EY) - (Y_j - EY)\right],\right.$$
$$\left.\left[(U_k - EU) - (U_l - EU)\right]\left[(V_k - EV) - (V_l - EV)\right]\right\}. \tag{12}$$

There are $2n^3 - n^2$, $n(n-1)(n-2)(n-3)$, $2n(n-1)$ and $4n(n-1)(n-2)$ items for the four conditions respectively. We can further separate (12) into 16 components and calculate their values under different conditions. The derivation is straightforward. Our conclusion thus holds. □

Theorem 3. *The sample mean of $(X - EX)(Y - EY)(U - EU)(V - EV)$ is an asympototically unbiased estimator of δ_4,*

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)\left(Y_i - \frac{1}{n}\sum_{j=1}^{n}Y_j\right)\right.$$
$$\left.\left(U_i - \frac{1}{n}\sum_{j=1}^{n}U_j\right)\left(V_i - \frac{1}{n}\sum_{j=1}^{n}V_j\right)\right]$$
$$= \left(1 - \frac{4}{n} + \frac{6}{n^2} - \frac{3}{n^3}\right)\delta_4 + \left[\frac{2(n-1)}{n^2} - \frac{3(n-1)}{n^3}\right](\delta_2 + \sigma_2 + \tau_2) \xrightarrow{n\to\infty} \delta_4. \tag{13}$$

Proof 3. *Equation (13) can be rewritten as*

$$\sum_{i=1}^{n} E\left\{\left[(X_i - EX) - \left(\frac{1}{n}\sum_{j=1}^{n}X_j - EX\right)\right]\right.$$
$$\left[(Y_i - EY) - (\frac{1}{n}\sum_{j=1}^{n}Y_j - EY)\right]$$
$$\left[(U_i - EU) - \left(\frac{1}{n}\sum_{j=1}^{n}U_j - EU\right)\right] \tag{14}$$
$$\left.\left[(V_i - EV) - \left(\frac{1}{n}\sum_{j=1}^{n}V_j - EV\right)\right]\right\}.$$

Again (14) can be separated into 16 components which are solvable under the independence assumption. The rest of the proof is omitted due to page limit. □

By integrating (8-13), the covariance of the LD contrast test statistics can be estimated. Note that the variance of the standardized LD contrast test statistic is approximately 1,

$$Var(T_{XY}) = Var\left[\frac{\Delta\hat{LD}_{XY}}{\sqrt{Var(\Delta\hat{LD}_{XY})}}\right] \approx \frac{Var(\Delta\hat{LD}_{XY})}{Var(\Delta\hat{LD}_{XY})} = 1. \tag{15}$$

Therefore, the covariance of T_{XY} and T_{UV} can be reduced to the corresponding correlation coefficients,

$$corr(T_{XY}, T_{UV}) \approx cov(T_{XY}, T_{UV}). \tag{16}$$

The test statistic for region-based interactions

To aggregate SNP-SNP interaction test statistics, a minimum p-value based method is adopted. In detail, we assume a multivariate normal distribution $MVN(0, \Sigma)$ for the observed test statistics $z_i, i = 1, 2, ..., k_1k_2$, where k_1 and k_2 are the number of SNPs in the two regions. The covariance matrix Σ is estimated using (8-13).

Then the region-based p-value is defined as the probability that we observe a value that is larger than the largest absolute value of SNP-SNP interaction test statistics under $MVN(0, \Sigma)$. Denote the absolute value of the test statistic related to the minimum p-value as T:

$$T = \left|\Phi^{-1}\left(\frac{min(p_i,\ i = 1, 2, .., k_1k_2}{2}\right)\right|. \tag{17}$$

Then the p-value for this region-region interaction reads,

$$\begin{aligned} p_{\text{region-region}} &= Pr\left[max(|z_i|,\ i = 1, 2, .., k_1k_2)\right. \\ &\geq \left. T \mid z_i \sim MVN(0, \Sigma)\right] \\ &= 1 - Pr\left[max(|z_i|,\ i = 1, 2, .., k_1k_2)\right. \\ &< \left. T \mid z_i \sim MVN(0, \Sigma)\right] \\ &= 1 - Pr\left[\{|z_i| < T,\ i = 1, 2, .., k_1k_2\} \mid z_i\right. \\ &\sim \left. MVN(0, \Sigma)\right]. \end{aligned} \tag{18}$$

In this paper, We use the results of GBOOST [18], the GPU version of BOOST, to specify candidate regions. The regions could also be selected by checking potential pathogenic pathways or protein-protein interaction networks.

Results

We conducted simulations under various settings to examine whether the proposed method can correctly control type-I-error rates and outperform SNP-based methods in terms of statistical power. To mimic real LD patterns, we picked all genotyped SNPs from two genomic regions (A and B) with intensive LD patterns in the dataset from Myocardial Infarction Genetics Consortium (MIGen) [19]. Region A is of size 157.874 kbp, located in chromosome 1, with 34 genotyped SNPs inside and 9 tag SNPs selected by haploview. Region B is of size 267.528 kbp, located in chromosome 3, with 50 genotyped SNPs and 10 tag SNPs.

We developed the software RRIntCC in C++. The source code of RRIntCC is available at http://bioinformatics.ust.hk/RRIntCC.html. The results of RRIntCC and SNP-based methods were compared for empirical power experiments. We further applied

RRIntCC to MIGen and a renal complication dataset of type 2 diabetes (T2D) patients. RRIntCC reported several significant region pairs in both datasets while conventional SNP-based interaction detection tools failed to identify any SNP pairs.

Type-I-Error rate control

For type-I-error rate evaluation, we randomly selected 1000, 2000, 3000, 4000, and 5000 samples from MIGen dataset and maintained their genotype values to preserve the LD patterns. Phenotype values for the randomly picked samples were assigned using a Bernoulli distribution with equal probabilities for case and control disease status. 1000 simulations were run for each sample size to determine the empirical type-I-error rates under two commonly used significance levels, i.e. 0.05 and 0.01. We repeat the experiment 20 times to examine the robustness of empirical type-I-error rates. As shown in Fig 1, simulations of empirical type-I-error rates indicated that the results of RRIntCC are not inflated at given significance levels.

Empirical statistical power

For power evaluation, phenotype values were generated using the public software GWASimulator [20], which uses haplotype information to simulate LD structure and produces phenotype values according to preset disease prevalence, causal SNPs and interactions with certain effect sizes. In total, 12084 haplotypes of these two regions were generated by PLINK [7]. We performed 1000 simulations for 1000, 2000, 3000, 4000, and 5000 samples, respectively. Results of original LD contrast test (LDCont) and GBOOST were also given for comparison.

GWASimulator simulated genotypes of all SNPs in the two regions, while only the tag SNPs were analyzed. Even though non-tag SNPs could be selected as causal SNPs, we can still observe interaction effects between tag SNPs due to LD between tag SNPs and non-tag SNPs. We designed six experimental settings with different tag status and allele frequencies for the causal interacted SNP pair. The effect sizes were determined by the relative risk ratio. The increment of relative risk ratio by observing one disease allele was set as $\sqrt{2}$, so that the ratios for genotype combinations 1/1, 1/2, 2/1, and 2/2 were 2, $2\sqrt{2}$, $2\sqrt{2}$, and 4, respectively. The results are summarized in Table 1. Under all settings, RRIntCC achieves a higher power than LDCont. GBOOST outperforms RRIntCC and LDCont when the MAFs of both causal SNPs are large. However, when the MAF of even one causal SNP goes down, the power of GBOOST drops dramatically and RRIntCC is the most powerful method under such settings. Even in the cases where both MAFs are large, RRIntCC is still valuable when sample size is small. The results support the use of our region-based interaction detection method in GWAS studies, especially considering that GWAS datasets usually have quite limited sample sizes compared to the huge number of SNPs.

Experiment using real datasets

We applied our method to the dataset of Myocardial Infarction Genetics Consortium (MIGen) with 649370 genotyped SNPs and 2967/3075 cases/controls, and the renal complication dataset collected in Hong Kong with 1257031 SNPs and 882/2231 cases/controls. Current computation capability cannot support whole-genome interaction analysis using LD contrast test. Instead, GBOOST [18] was first used as probes to generate region-pairs qfor region-based interaction analysis. We adopted 5×10^{-10} as a suggestive p-value threshold to screen out SNP pairs

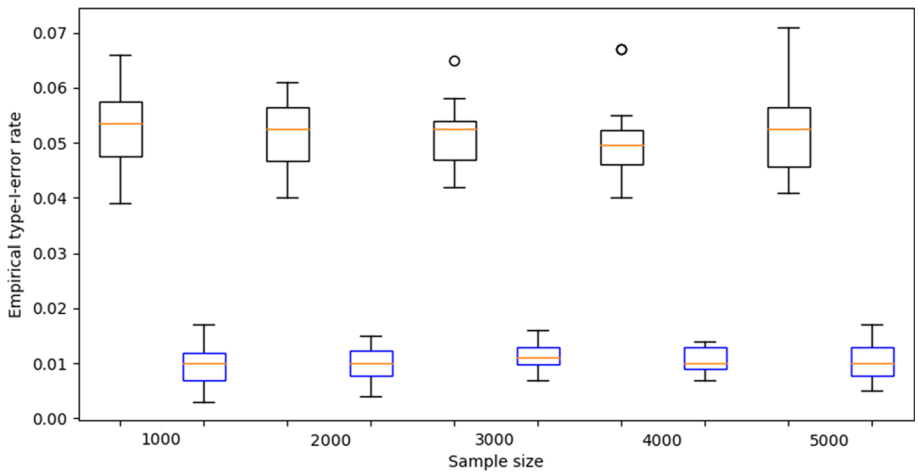

Fig. 1 The boxplots of empirical type-I-error rates at the significant levels of 0.05 (black) and 0.01 (blue)

Table 1 Empirical statistical power results

		1000	2000	3000	4000	5000
19(0.424) ∼28(0.414)	RRIntCC	**0.247**	0.564	0.806	0.913	0.975
	LDCont	0.205	0.524	0.778	0.888	0.968
	GBOOST	0.240	**0.624**	**0.872**	**0.961**	**0.994**
19(0.424) ∼22*(0.413)	RRIntCC	**0.255**	0.545	0.814	0.924	0.979
	LDCont	0.214	0.489	0.793	0.905	0.969
	GBOOST	0.218	**0.609**	**0.885**	**0.968**	**0.998**
15(0.067) ∼22*(0.413)	RRIntCC	**0.244**	**0.548**	**0.772**	**0.880**	**0.964**
	LDCont	0.188	0.496	0.724	0.849	0.953
	GBOOST	0.058	0.211	0.411	0.559	0.731
23*(0.067) ∼22*(0.413)	RRIntCC	**0.307**	**0.631**	**0.882**	**0.954**	**0.986**
	LDCont	0.264	0.574	0.856	0.942	0.975
	GBOOST	0.088	0.272	0.548	0.713	0.838
15(0.067) ∼25(0.094)	RRIntCC	**0.116**	**0.266**	**0.398**	**0.551**	**0.667**
	LDCont	0.072	0.204	0.323	0.480	0.612
	GBOOST	0.012	0.060	0.108	0.224	0.285
23*(0.067) ∼25(0.094)	RRIntCC	**0.110**	**0.282**	**0.502**	**0.638**	**0.790**
	LDCont	0.081	0.220	0.428	0.576	0.729
	GBOOST	0.025	0.064	0.161	0.259	0.397

The indices are the order of SNPs in their corresponding regions, * means this SNP is a tag SNP, and the values in the brackets denote minor allele frequencies (MAFs).

that are unlikely to be associated. The remaining SNP pairs were clumped into regions with size 200 kbp, which is roughly the size of typical genes. After identifying the ranges of clumped regions, all genotyped SNPs in MIGen dataset were mapped into these regions. For computation efficiency, the maxmium number of SNPs in each region was set to be 31, so that the total number of SNP-SNP interactions within each region pair was controlled below 1000. The choice of this number is arbitrary. In case that the real number of SNPs inside a region is larger than this limit, we randomly choose 31 SNPs to represent this region.

Table 2 lists the top four SNP pairs found by GBOOST in the MIGen dataset and their corrected family-wise error rates (cFWER). None of them can pass the Bonferroni-corrected p-value threshold. Moreover, even the smallest p-value is 100 times larger than the threshold. Table 3 lists the top four region pairs found by RRIntCC. One region pair, chr3: [177577480, 177777480] ∼ chr7: [81695481, 81895481], passes the Bonferroni-corrected p-value threshold. The second and third region pairs share the same region in chr3 and overlap in the region in chr20, which indicates that these two region pairs actually refer to only one region pair with size larger than the preset 200 kbp. Therefore, we further analyze the region interaction between chr3: [187498383, 187698383] with size 200 kbp and chr20: [39109460, 39444799] with size 335.339 kbp, leading to a cFWER of 0.0536. Multiple genes, including CACNA2D1, DGKG, AK057298, TOP1, BC035080, PLCG1, ZHX3, LPIN3, and EMILIN3, are located in these two region pairs. CACNA2D1 has been found to be involved in cardiomyopathy pathway [21, 22]. Besides, ZHX3 is reported to be associated with left ventricle wall thickness [23]. Both ZH3 and EMILIN3 are reported to be associated with resting heart rate [24]. The regions identified by RRIntCC might provide clues for factors affecting myocardial infarction risks.

We also applied GBOOST and RRIntCC to the renal complication dataset. GBOOST has no significant finding, while RRIntCC found one region pair, chr12: [103040398, 103240398] and chr15: [33102602, 33302602], with a cFWER of 0.00382. Two genes, PAH and FMN1, are involved in this region pair. Both PAH and FMN1 were reported to be related to kidney disorders

Table 2 Top four SNP pairs found by GBOOST in the MIGen dataset

SNP pairs	p-value	cFWER
rs4678428 (chr3) ∼ rs9961565 (chr18)	2.588×10^{-11}	> 1
rs17626606 (chr5) ∼ rs11190346 (chr10)	2.679×10^{-11}	> 1
rs11925209 (chr3) ∼ rs1501909 (chr5)	3.006×10^{-11}	> 1
rs6930292 (chr6) ∼ rs114313 (chr6)	3.026×10^{-11}	> 1

Table 3 Top four region pairs found by RRIntCC in the MIGen dataset

region pairs	*p*-value	cFWER
chr3: [177577480, 177777480] ∼ chr7: [81695481, 81895481]	1.652×10^{-10}	0.0186
chr3: [187498383, 187698383] ∼ chr20: [39244799, 39444799]	5.363×10^{-10}	0.0603
chr3: [187498383, 187698383] ∼ chr20: [39109460, 39309460]	7.497×10^{-10}	0.0843
chr2: [184236258, 184436258] ∼ chr13: [29010198, 29210198]	7.835×10^{-9}	0.8814

[25][26], which implies a potentially target pathway for the study of renal complications in patients with T2D.

Discussion

There still remain several issues that could be improved in our method. First, the computation complexity of calculating the covariance matrix is $O(n^2)$, which is unacceptable for whole genome analysis. Second, the genomic resolution has been sacrificed by replacing SNPs with regions. One potential remedy is to extend statistical fine mapping methods for interaction detection to determine the leading SNP pairs within the significant region pairs.

Conclusions

In this paper, we proposed a region-based interaction detection method named RRIntCC. We derived the correlation coefficients between SNP-SNP interaction test statistics by using LD contrast test. We aggregated SNP-SNP interaction test statistics by assuming a multi-variate normal distribution with the estimated covariance matrix to account for the potential intensive LD pattern within the regions. By using region-based strategy, we reduced the total number of tests and were therefore able to use a less stringent Bonferroni-corrected *p*-value threshold. Simulation results support that our region-based strategy outperforms SNP-based method in terms of statistical power at similar type-I-error rates.

Abbreviations

bp: Base Pair; cFWER: Corrected family-wise error rate; CLD: Composite linkage disequilibrium; GWAS: Genome-wide association study; LD: Linkage disequilibrium; MAF: Minor allele frequency; MIGen: Myocardial infarction genetics consortium; RRIntCC: Region-region interaction detection for case-control studies; SNP: Single nucleotide polymorphism; T2D: Type II diabetes

Acknowledgements

We would like to thank the anonymous reviewers for their helpful suggestions.

About this supplement

This article has been published as part of BMC Medical Genomics, Volume 12 Supplement 7, 2019: Selected articles from the 14th International Symposium on Bioinformatics Research and Applications (ISBRA-18): medical genomics. The full contents of the supplement are available at url=https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-12-supplement-7

Authors' contributions

SZ designed the method, implemented the C++ package, performed statistical experiments and drafted the manuscript. WJ contributed to the design of the study and the design of the method. RCWM provided the data and contributed to the design of the study. WY conceived and supervised the study, and helped to draft the manuscript. All authors read and approved the final manuscript.

Funding

This study is partially funded by the Theme-based Research Scheme (project T12-402/13N) of the Hong Kong Research Grant Council (RGC). We like to declare that RGC is not involved in the design of the study, collection, analysis, and interpretation of data, as well as in the writing of the manuscript.

Availability of data and materials

The data of Myocardial Infarction Genetics Consortium are publicly available from The database of Genotypes and Phenotypes (dbGaP), accession number: phs000294.v1.p1. The data of renal complication in T2D patients are are available from the Theme-based Research Scheme (T12-402/13N) of the Hong Kong Research Grant Council (RGC) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Theme-based Research Scheme (T12-402/13N).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology,, Kowloon, Hong Kong, China. [2]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Kowloon, Hong Kong, China. [3]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Shatin, Hong Kong, China.

Received: 27 August 2019 Accepted: 10 September 2019
Published: 30 December 2019

References

1. Welter D, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42(D1):D1001–6.
2. Burton PR, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78.
3. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Human Genet. 2012;90(1):7—24.
4. Cordell HJ. Detecting gene–gene interactions that underlie human diseases. Nature Rev Genet. 2009;10(6):392–404.
5. Wan X, et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Human Genet. 2010;87(3):325–40.
6. Hu JK, Wang X, Wang P. Testing gene–gene interactions in genome wide association studies. Genet Epidemiol. 2014;38(2):123–34.
7. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Human Genet. 2007;81(3):559–75.
8. Ritchie MD, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Human Genet. 2001;69(1):138–147.
9. Moore JH, White BC. Springer: Machine Learning and Data Mining in Bioinformatics; 2007, pp. 166–75.
10. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. Nature Genet. 2007;39(9):1167–73.

11. Zaykin DV, Meng Z, Ehm MG. Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. Am J Human Genet. 2006;78(5):737–46.
12. Li MX, Gui HS, Kwan JS, Sham PC. GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am J Human Genet. 2011;88(3):283–93.
13. Liu JZ, et al. A versatile gene-based test for genome-wide association studies. Am J Human Genet. 2010;87(1):139–45.
14. Ma L, Clark AG, Keinan A. Gene-based testing of interactions in association studies of quantitative traits. PLoS Genet. 2013;9(2):e1003321.
15. Hill WG. Estimation of linkage disequilibrium in randomly mating populations. Heredity. 1974;33(2):229–39.
16. Wu X, Jin L, Xiong M. Composite measure of linkage disequilibrium for testing interaction between unlinked loci. Eur J Human Genet. 2008;16(5): 644–51.
17. Schaid DJ. Linkage disequilibrium testing when linkage phase is unknown. Genetics. 2004;166(1):505–12.
18. Yung LS, Yang C, Wan X, Yu W. GBOOST: a GPU-based tool for detecting gene–gene interactions in genome–wide case control studies. Bioinformatics. 2011;27(9):1309–10.
19. Kathiresan S, et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nature Genet. 2009;41(3):334–41.
20. Li C, Li M. GWAsimulator: a rapid whole-genome simulation program. Bioinformatics. 2007;24(1):140–2.
21. Cataldo S, Annoni GA, Marziliano N. The perfect storm?Histiocytoid cardiomyopathy and compound CACNA2D1 and RANGRF mutation in a baby. Cardiol Young. 2015;25(1):174–6.
22. Bourdin B, et al. Functional characterization of CaV$\alpha 2\delta$ mutations associated with sudden cardiac death. J Biol Chem. 2015;290(5):2854–69.
23. Wild PS, et al. Large-scale genome-wide analysis identifies genetic variants associated with cardiac structure and function. J Clin Investigation. 2017;127(5):1798–812.
24. Eppinga RN, et al. Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. Nature Genet. 2016;48(12):1557.
25. Lichter-Konecki U, Hipke CM, Konecki DS. Human phenylalanine hydroxylase gene expression in kidney and other nonhepatic tissues. Mole Genet Metabol. 1999;67(4):308–16.
26. Dimitrov BI, et al. Genomic rearrangements of the GREM1-FMN1 locus cause oligosyndactyly, radio-ulnar synostosis, hearing loss, renal defects syndrome and Cenani–Lenz-like non-syndromic oligosyndactyly. J Med Genet. 2010;47(8):569–74.

Publisher's Note