Measure gene - gene dependence using Kullback-Leibler divergence

(August 30, 2018)

Guanjie Chen [1], Ao Yuan [2], Tao Cai [3], Chuanming Li [4], Amy R. Bentley [1], Adebowale Adeyemo [1], Charles Rotimi [1]

[1]Center for Research on Genomics and Global Health, NHGRI, NIH, Bethesda, Maryland, 20892 USA

[2]Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC, 20057 USA

[3]Experimental Medicine Section, Laboratory of Sensory Biology, NIDCR, NIH, Bethesda, Maryland, 20892 USA

[4]Division of Scientific programs, National Institute on deafness and other communication disorders, Rockville, Maryland, 20852 USA

Abstract

Genome wide association studies (GWAS) are used to investigate genetic variants contributing to complex traits. Despite discovering hundreds of well replicated loci, large proportion of traits heritability estimates remain unexplained leading to the so called "missing heritability". Accounting for gene-gene interactions in disease gene mapping may help explain some of the"missing" heritability. At the level of the gene, interaction could be evaluated by accessing gene-gene dependence. Here, we propose a Kullback-Leibler type statistic for the analysis of gene-gene dependence, for uncorrelated SNPs in two genes. The Kullback-Leibler statistics is an asymptotic positive normal distribution under the null hypothesis of no relationship

between the two genes (SNPs), and normal under the alternative. Kullback-Leibler is unlike traditional parametric statistical methods such as linear and logistic regression that use multifactor dimensionality reduction (MDR) to address sparseness of data in high dimensions. The performance of the proposed method is evaluated by simulation studies with promising results. Applying proposed method to analyze real data, we identified gene-gene dependences among *RAB3A*, *MADD*, and *PTPRN* in the context of type 2 diabetes (T2D).

Keywords: Case-control population, Kullback-Leibler statistic, hypothesis test, gene-gene dependences, SNPs.

1. Introduction.

GWAS have been widely used to detect genetic factors influencing common complex diseases, such as Type 2 diabetes (T2D). To date, over 100 common genetic risk variants have been identified associated with T2D. However, the joint effects of these variants account for less than 10% of the heritability for T2D (Morris 2012). The gene-gene interactions may help explain the "missing" heritability.

There are extensive statistical methods for identifying gene-gene interactions, including: using false discovery rate (Onay et al. 2006), evaluating interactions between SNPs at multiple loci (Anno et al. 2008), a principal components method (Li et al. 2009), ensemble filtering approach (Yang et al. 2011), Cox model with censored responses (Lee et al. 2012), machine learning (Koo et al. 2013) for SNP-SNP interaction for cancer patients (Lin et al. 2013), a fast algorithm (Li et al. 2014), regression method (Park and Hastie, 2008), a linear model

(Aschard et al. 2015), gene-gene interaction for twin data (Buil et al. 2015), least trimmed square regression method (Kim and Park 2015), logistic regression (Kooperberg and Ruczinski, 2005)), nonparametric logistic model and use a basis function method to model gene effect curve and gene-gene interaction curve (Zhao et al. 2016), multi-dimensional reduction method (Moore, 2004; Chung et al., 2007), Bayesian method (Ferreira et al., 2007), Entropy method (Dong et al., 2008), gene-gene interaction for nuclear families (Martin et al, 2006), combinatorial method (Lou et al., 2008), neural network method (Motsinger et al., 2006), high order gene-gene interaction (Tsai et al., 2007), log-linear model (Lee et al, 2007), and a rare variants method (Zhu et al. 2010 and Yuan et al. 2012). Cordell (2009) and Gilbert-Diamond and Moore (2011) gave nice reviews on methods in this topic.

Most existing work on gene-gene interactions use a regression model, requiring a response variable, genes and other covariates. The interaction is represented by the gene by gene cross effects of the regression parameters. These method only apply to a small number of SNPs at each gene, so these methods are for local interactions. The genetic functional model, such as Fan et al. (2015) and Xu et al. (2017) and citations there in, can handle large number of genes in a single model. Here we consider a different framework using the SNP data from two genes to evaluate dependence and test dependence difference between cases and controls. Our method investigate the global dependences between two sets of SNPs from two genes, it tells us if the two genes have dependences or independences, not to pinpoint the specific SNPs for interaction (SNP-level interaction). Thus the mentioned methods are not applicable to this case.

After conducting linkage disequilibrium (LD) pruning to obtain a set of independent SNPs within each gene, we stratify by case-control status evaluate the dependence of SNPs in one gene with SNPs in the other gene by Kullback-Leibler statistic. Here we construct a Kullback-Leibler statistic which is asymptotic positive normal. Its cut-off point and other related quantities can be evaluated in a closed form, making it more practical to use. The method can be used to analyze the global gene-gene relationship, and evaluate the gene-gene dependence structure difference between the cases and controls. The proposed method is easy to use and can screen the genome wide level, and give a general guide for which pairs of genes are related as a whole, which pairs are not. Thus may be used as a first step analysis of gene-gene interactions in genome-wide level. After some pairs of genes are identified as related, then existing local methods can be used to further identify specific genes which are responsible for such interactions.

Simulation studies are conducted to evaluate the performance of the proposed method and showed promising results. Then the method is applied to a real data to evaluate dependences between variants in genes known to encode proteins that form a complex (*PTPRN*, *MADD* and *RAB3*) which regulates insulin release, and difference dependences between T2D cases and controls in the Africa America Diabetes Mellitus (AADM) study.

In Section 2, we describe the background and the problem, and introduce the proposed method. We considered two cases: the gene-gene dependence between given two genes (two sets of SNPs) for one population; and the gene-gene dependence difference between cases and controls. The asymptotic distributions of the test statistics are studied. In Section 3, we conduct simulation studies to evaluate

the performance of the proposed method; Section 4 gives the analysis of real data using the proposed method, and Section 5 give a brief discussion.

2. The Method. Our method uses discrete trait data with genotyped SNPs from unrelated individuals. As there are large number of SNPs in the genome such as SNP array data, it is useful to generate a pruned subset of SNPs that are in approximate linkage equilbrium with each other. So before the gene-gene dependence analysis, linkage disequilibrium based on SNP pruning was used to select indepedent SNPs in each gene. Let $(\mathbf{g}_{i1}, \mathbf{g}_{i2})$ $(i = 1, ..., n)$ be the observed genotypes from two genes of n unrelated individuals, $\mathbf{g}_{i1}$ is a sequence of m_1 unrelated SNPs in gene 1, $\mathbf{g}_{i2}$ a sequence of m_2 unrelated SNPs in gene 2. let $m = m_1 + m_2$ with m_1 and m_2 in the hundreds to thousands. Each SNP can have one of three genotypes, we arbitrarily code them as 1,2, and 3. The position relationships among components of $\mathbf{g}_{i1}$ and $\mathbf{g}_{i2}$ are known. As there are large numbers of SNPs in $\mathbf{g}_{i1}$ and $\mathbf{g}_{i2}$, we can review them as functions $g_{i1}(s), s \in S$ and $g_{i2}(t), t \in T$, with S and T the index set for gene 1 and gene 2 respectively. The $g_{i1}(\cdot)$'s and $g_{i2}(\cdot)$'s are independent and identically distributed random realizations of some (unknown) $\{1, 2, 3\}$ stochastic processes $g_1(\cdot)$ on S and $g_2(\cdot)$ on T.

We consider two cases: the gene-gene dependence between a given pair of sets of SNPs for one population, within this case we also distinguish two scenarios, a) gene-gene dependences between a given pair of sets of SNPs. b) gene-gene dependences among two sets of SNPs from two genes; and the difference of gene-gene dependence patterns in cases and controls.

2.1 Gene-gene dependence with one population. At each fixed point $s \in S$, let $p_k(s) = P\big(g_1(s) = k\big)$ be the population frequency of SNP type k at position s for gene 1 ($k = 1, 2, 3$), $q_k(t) = P\big(g_2(t) = k\big)$ be that at position $t \in T$ for gene 2, and $P_{jk}(s,t) = P\big(g_1(s) = j; g_2(t) = k\big)$ be that of the composite SNP type (j,k) at position s of gene 1 and position t of gene 2 ($j, k = 1, 2, 3$). These are unknown functions to be estimated. Let $I(\cdot)$ be the indicator function. These quantities are estimated by

$$\hat{p}_k(s) = \frac{1}{n}\sum_{i=1}^{n} I(g_{i1}(s) = k), \qquad (k = 1, 2, 3; s \in S),$$

$$\hat{q}_k(t) = \frac{1}{n}\sum_{i=1}^{n} I(g_{i2}(t) = k), \qquad (k = 1, 2, 3; t \in T) \tag{1}$$

and

$$\hat{P}_{jk}(s,t) = \frac{1}{n}\sum_{i=1}^{n} I(g_{i1}(s) = j; g_{i2}(t) = k), \qquad (k, j = 1, 2, 3; s \in S; t \in T).$$

If the two sets of unrelated SNPs have no interactions, they will be independent across all positions, while lack of independence indicates the presence of an interaction. The problem of whether there are gene-gene interactions in the two sets of SNPs can be formulated, then, as a statistical test of the null hypothesis $H_0 : P_{jk}(s,t) = p_j(s)q_k(t) \quad \text{for all} \quad (j, k = 1, 2, 3; s \in S; t \in T)$ vs. the alternative hypothesis $H_1 : P_{jk}(s,t) \neq p_j(s)q_k(t)$, for some $(j, k = 1, 2, 3; s \in S; t \in T)$.

Since the hypothesis involves a function of values for $(j, k = 1, 2, 3; s \in S; t \in T)$, typically the appropriate test statistic has an asymptotic mixture chi-squared distribution, for which the cut-off point and other related values cannot be obtained in closed form, and a numerical method is needed to compute them, which is not convenient in practice. Here, we develop a special construction for the test

statistic, so that its asymptotic distribution is positive normal, to be defined later. All quantities of interest under this distribution have a closed form, making this statistic more practical to use.

To test H_0 vs H_1, let

$$D = D(P||pq) = \sum_{j=1}^{3}\sum_{k=1}^{3}\int_S\int_T P_{jk}(s,t)\log\frac{P_{jk}(s,t)}{p_j(s)q_k(t)}dsdt. \tag{2}$$

be the Kullback-Leibler divergence between P and pq. Here we adopt the convention $0\log 0 = 0$, and use integration for summation over $S \times T$ to distinguish the summation over (j,k). It is known that $D(P||pq) \geq 0$, with $D(P||pq) = 0$ if and only if H_0 holds. Thus, if there is no interaction between $g_1(\cdot)$ and $g_2(\cdot)$, and H_0 holds, $D(P||pq) = 0$. On the other hand, if $g_1(\cdot)$ and $g_2(\cdot)$ are perfectly correlated: $g_1(s) = k$ if and only if $g_2(s) = k$ for $k = 1, 2, 3$ and $S = T$, then $P_{jk}(s,t) = p_j(s)$ and $D(P||pq) = -\sum_{j=1}^{3}\int p_j(s)\log p_j(s)ds = H(p)$, the entropy of p.

The estimator of $D(P||p,q)$ is

$$\hat{D}_n = \hat{D}_n(P||pq) = \sum_{j=1}^{3}\sum_{k=1}^{3}\int_S\int_T \hat{P}_{jk}(s,t)\log\frac{\hat{P}_{jk}(s,t)}{\hat{p}_j(s)\hat{q}_k(t)}dsdt. \tag{3}$$

The followinng results characterize the asymptotic properties of $\hat{D}_n$.

Theorem 1. *As $n \to \infty$, $\hat{D}_n \to D$, a.s.*

Theorem 2. *As $n \to \infty$,*

i) under H_0,

$$\sqrt{n}\hat{D}_n \xrightarrow{D} C_0 + N(0, \sigma_0^2),$$

where $C_0 = \lim_n \sqrt{n} E_{H_0}(\hat{D}_n)$ *and* σ_0^2 *is given at the end of the proof.*

ii) Under H_1,

$$\sqrt{n}(\hat{D}_n - D) \xrightarrow{D} N(0, \sigma_1^2),$$

σ_1^2 *is given at the end of the proof.*

Let $\Phi^{-1}(\cdot)$ be their quantile function of $N(0,1)$. For given nominal level α, H_0 is rejected if $T_n := \sqrt{n}\sigma_0^{-1}\hat{D}_n > C_1 + \Phi^{-1}(1-\alpha)$, with $C_1 = C_0\sigma_0^{-1}$.

Power. The power of the test at D is

$$\beta(D) = P_{H_1}\Big(\sqrt{n}\sigma_0^{-1}\hat{D}_n > C_1 + \Phi^{-1}(1-\alpha)\Big)$$

$$= P_{H_1}\Big(\sqrt{n}\sigma_1^{-1}(\hat{D}_n - D) > \sigma_1^{-1}C_1 + \sigma_1^{-1}\Phi^{-1}(1-\alpha) - \sqrt{n}\sigma_1^{-1}D\Big)$$

$$\approx 1 - \Phi\Big(\sigma_1^{-1}C_1 + \sigma_1^{-1}\Phi^{-1}(1-\alpha) - \sqrt{n}\sigma_1^{-1}D\Big).$$

2.2 Dependences within a set of unrelated SNPs. In the above we discussed dependences between two sets of unrelated SNPs, here we investigate dependences within a set of unrelated SNPs. We concentrate on the data $\{\mathbf{g}_{i1} : i = 1, ..., n\}$ in a set of unrelated SNPs. In this case we denote $P_{jk}(s,t) = P(g_1(s) = j; g_1(t) = k)$, and

$$\hat{P}_{jk}(s,t) = \frac{1}{n}\sum_{i=1}^{n} I(g_{i1}(s) = j; g_{i1}(t) = k).$$

We are interested in testing the null hypothesis $H_0 : P_{jk}(s,t) = p_j(s)p_k(t)$ for all $1 \le j,k \le 3$ and $(s,t) \in S^2$ with $s \ne t$, vs. the alternative $H_1 : P_{jk}(s,t) \ne p_j(s)p_k(t)$ for some $1 \le j,k \le 3$ and $(s,t) \in S^2$ with $s \ne t$. As before, let

$$D = D(P||pp) = \sum_{j=1}^{3}\sum_{k=1}^{3}\int_{S:s\ne t}\int_{S} P_{jk}(s,t)\log\frac{P_{jk}(s,t)}{p_j(s)p_k(t)}dsdt \qquad (4)$$

be the Kullback-Leibler divergence between P and pp. Here we adopt the convention $0 \log 0 = 0$. It is known that $D(P||pp) \geq 0$, with $D(P||pp) = 0$ if and only if H_0 holds. Thus, if there is no dependence within $g_1(\cdot)$, and H_0 holds, $D(P||pp) = 0$. However, if $g_1(\cdot)$ is perfectly correlated: $g_1(s) = k$ if and only if $g_1(t) = k$ for $k = 1, 2, 3$, then $P_{kk}(s,t) = p_k(s)$ and $D(P||pq) = -\sum_{j=1}^{3} \int p_j(s) \log p_j(s) ds = H(p)$, the entropy of p.

The estimator of $D(P||p,p)$ is

$$\hat{D}_n = \hat{D}_n(P||pp) = \sum_{j=1}^{3} \sum_{k=1}^{3} \int_{S:s \neq t} \int_S \hat{P}_{jk}(s,t) \log \frac{\hat{P}_{jk}(s,t)}{\hat{p}_j(s)\hat{p}_k(t)} ds dt. \tag{3}$$

We have

Corollary. *i). As $n \to \infty$, $\hat{D}_n \to D$, a.s.*

ii). As $n \to \infty$, then under H_0,

$$\sqrt{n}\hat{D}_n \xrightarrow{D} C_0 + N(0, \sigma_0^2),$$

where $C_0 = \lim_n \sqrt{n} E_{H_0}(\hat{D}_n)$ and σ_0^2 is given in the proof.

iii) As $n \to \infty$, under H_1,

$$\sqrt{n}(\hat{D}_n - D) \xrightarrow{D} N(0, \sigma_1^2),$$

σ_1^2 is given in the proof.

The rejection rule of H_0 and the power computation are similar to testing interactions across genes.

2.3 Gene-gene dependence difference between cases and controls.

In practice, often we have case-control data, and it is of interest to know if the dependences between cases and controlls are different or not. For this, suppose we have case data n on two genes and control data m on the same genes. Let D_1 and D_0 be the estimated Kullback-Leibler measure for the two populations. We can use (3) to compute the estimated Kullback-Leibler measure $\hat{D}_n$ and $\hat{D}_m$ for the two populations. We want to test $H_0 : D_1 = D_0$ vs $H_1 : D_1 \neq D_0$. For this, let

$$T_{n,m} = \sqrt{\frac{nm}{n+m}} \frac{\hat{D}_n - \hat{D}_m}{\left(\frac{n}{n+m}\hat{\sigma}_1^2 + \frac{m}{n+m}\hat{\sigma}_0^2\right)^{1/2}},$$

where $\hat{\sigma}_1^2$ and $\hat{\sigma}_0^2$ are the estimated variances for $\hat{D}_n$ and $\hat{D}_m$ as given in Theorem 2 ii). Then as $\min\{n, m\} \to \infty$, under H_0, $T_{n,m} \sim N(0, 1)$, and the α rejection rule for H_0 is $|T_{n,m}| > \Phi^{-1}(1 - \alpha/2)$.

3. Simulation Study.

To model the dependence of the two sets of unrelated SNPs, we use a continuous method. We take $S = T = [-5, 5]$, for some combinations of (m_1, m_2), $n = 5000$ and $N = 1000$ repetitions. Set $p_1(s) = [1 + sin(s/10)/2]/3$, $p_2(s) = [1 + sin(s/10)/3]/3$, $p_3(s) = 1 - p_1(s) - p_2(s)$; $q_1(s) = [1 + sin(s/20)/3]/3$, $q_2(s) = [1 + sin(s/20)/2]/3$, $q_3(s) = 1 - q_1(s) - q_2(s)$. We use D_0 to denote the average value of $\sqrt{n}(\hat{D}_n - D)$ (over repetitions), and $\hat{\sigma}_0^2$ for estimated σ_0^2.

Simulation Study 1: No dependence. For each subject, at gene position (s, t), with sample composite SNP type (j, k) and probability $P_{jk}(s, t) = p_j(s)q_k(t)$, $(j, k = 1, 2, 3)$, we generated 1,000 simulations. Type 1 error rates are presented in Table 1.

Table 1. Type 1 errors in Simulation study Case 1 - indepdendece study

Genes (m_1/m_2)	N	No. repeat	Type 1 Errors	D_0	$\hat{\sigma}_0^2$
10 / 10	1,000	1,000	4.60%	0.06329	0.00139
50 / 50	1,000	1,000	4.90%	0.00206	0.00142
100 / 100	1,000	1,000	4.70%	0.00197	0.00137
50 / 100	1,000	1,000	4.80%	0.00020	0.00140
50 / 200	1,000	1,000	4.90%	0.00203	0.00133
50 / 250	1,000	1,000	4.90%	0.00199	0.00134

Simulation Study 2: Dependence. For each subject, at gene position (s,t), and sample composite SNP type (j,k) with probability

$$P_{jk}(s,t) = \frac{p_j(s)q_k(t) + \lambda p_j(s)(1-q_k(t))}{\sum_{i=1}^{3}\sum_{r=1}^{3}(p_i(s)q_r(t) + \lambda p_i(s)(1-q_r(t))}, \qquad (j,k=1,2,3),$$

we perform 1,000 repetitions.

Here $0 \leq \lambda \leq 1$ control the amount of dependence, $\lambda = 0$ corresponds to independence; $\lambda = 1$ corresponds to perfect dependence with $P_{jk}(s,t) = p_j(s)/3$.

Table 2. Power calculation in Simulation study Case 2-dependence study

Genes (m_1/m_2)	N	No. repeat	Power	λ
10 /10	1,000	1,000	82.40%	0.0313
50 /50	1,000	1,000	80.80%	0.0022
100 /100	1,000	1,000	81.80%	0.0022
50 /100	1,000	1,000	82.20%	0.0022
50 /200	1,000	1,000	82.6%	0.0021
50 /250	1,000	1,000	82.2%	0.0021

Figure 1 presents the relationship between the total number of SNPs within the two genes,the estimated D_0 (red line) in simulation study 1 (no dependence), and the estimated λ in simulation study 2 (with dependence).

4. Application to real data

The Africa America Diabetes Mellitus (AADM) study is a large ongoing genetic epidemiology study of type 2 diabetes (T2D) (Rotimi et al 2001). Demographic information was collected using standardized questionnaires across the AADM study centers in Nigeria, Ghana, and Kenya. Anthropometric and other clinical parameters were measured by trained study staff during a clinic visit. The diagnosis of T2D was based on the 1999 American Diabetes Association Expert Committee criteria (Engelgau 2000): a fasting plasma glucose concentration $\geq$ 126 mg/dL (7.0 mmol/L), a 2-hour post load value in the oral glucose tolerance test $\geq$ 200 mg/dL (11.1 mmol/L) on more than one occasion, or taking medication for physician-diagnosed T2D. Of the 1,808 African samples, 1,046 (57.85%) were T2D cases,

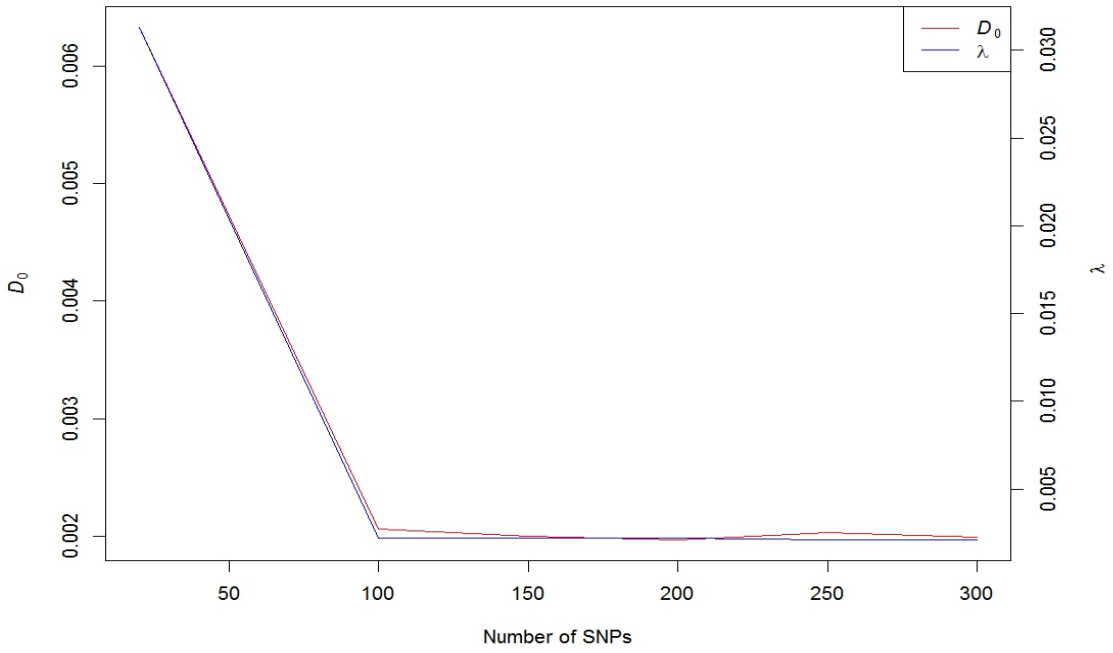

Figure 1: Relationships between Number of SNPs and estimated values of D_0 in simulation 1 and λ in simulation 2

and 762 (42.15%) were controls. Genotyping was performed at the Children's Hospital of Philadelphia core laboratory using the Affymetrix Axiom PanAFR Array. A total of 2,217,748 SNPs were subjected to quality control filters as follows: call rate $< 90\%$ (excluded 10,282), minor allele frequency (MAF) < 0.01 (excluded 45,562), and non-autosomal SNPs (excluded 61,087). Since no comparison gene-gene interaction method was used in genome wide level, we did not try to evaluate gene-gene dependences in genome wide level. To evaluate our method, we selected 3 genes that encode proteins with interrelated functions that affect pathways important to T2D risk. MADD acts as a guanosine triphosphate exchange protein for Rab3A (Coppola 2002) and Rab27a which play critical roles in glucose-stimulated insulin release from β-cells (Kasai 2005). MADD and Rab3A molecules are linked to dense core secretory vesicles (DCVs) by Ptprn/2 (phogrin), a transmembrane protein expressed in cells with stimulus-coupled peptide hormone secretion, including pancreatic β-cells. It is localized to the membrane of insulin-containing DCVs (Caromile 2010). Figure 2 represents the protein complex of Rab3, MADD, and Ptprn that stimulates DCV to release insulin. Here, we used AADM genotype data to test for pairwise gene-gene dependences among *RAB3A*, *MADD*, and *PTPRN*. SNPs within these genes was performed using human genome reference 19 (HG19). We conducted LD pruning (window size in 50 SNPs and calculate LD between each pair of SNPs in the window; remove one of a pair of SNPs if the LD is greater than 0.5; shift window 5 SNPs forward and repeat the procedure) in PLINK (Purcell 2007) to produce an independent set of SNPs for each gene in study (1,379 SNPs in PTPRN-PTPRN2, 15 SNPs in MADD, and 211 SNPs in RAB3A). The pair-wise gene-gene dependences results are presented in Table 3.

For comparison, we also tested for gene-gene interactions using an MDR method, which comprised 3 steps. First, association is performed for each of the multi-locus combinations of genotypes. Second, based on the p values and direction of β values resulting from association testing, the multi-locus genotypes can be classified as non-risk (p values ≥ 0.10), high risk (β value > 0 and p value < 0.10), or low risk (β value < 0 and p value < 0.10). To avoid inflation of parameter estimates, the multi-locus genotypes are combined by risk category. Finally, two new association tests are performed using the Wald statistic: high risk vs. low risk + no risk, and low risk vs. high risk + no risk groups. The test statistic for the epistatic effects will be the maximum of the two tests. We used a Bonferroni p value correction, based on the number of tests performed per gene pair, to set the threshold for statistical significance. Results using the MDR method are presented in Table 4.

Table 3. Results of Pairwise Gene-Gene Dependences in AADM Study

Gene 1 (Num. of SNPs)	Gene 2 (Num. of SNPs)	Z values	P values
PTPRN/2 (1,379)	*MADD* (15)	19.53	6.70×10^{-85}
PTPRN/2 (1,379)	*RAB3A/C* (211)	16.38	2.77×10^{-60}
RAB3A/C (211)	*MADD* (15)	15.99	1.30×10^{-57}

For the real data, we can use the formula to compute $\hat{\sigma_0^2}$, or use bootstrape to compute it.

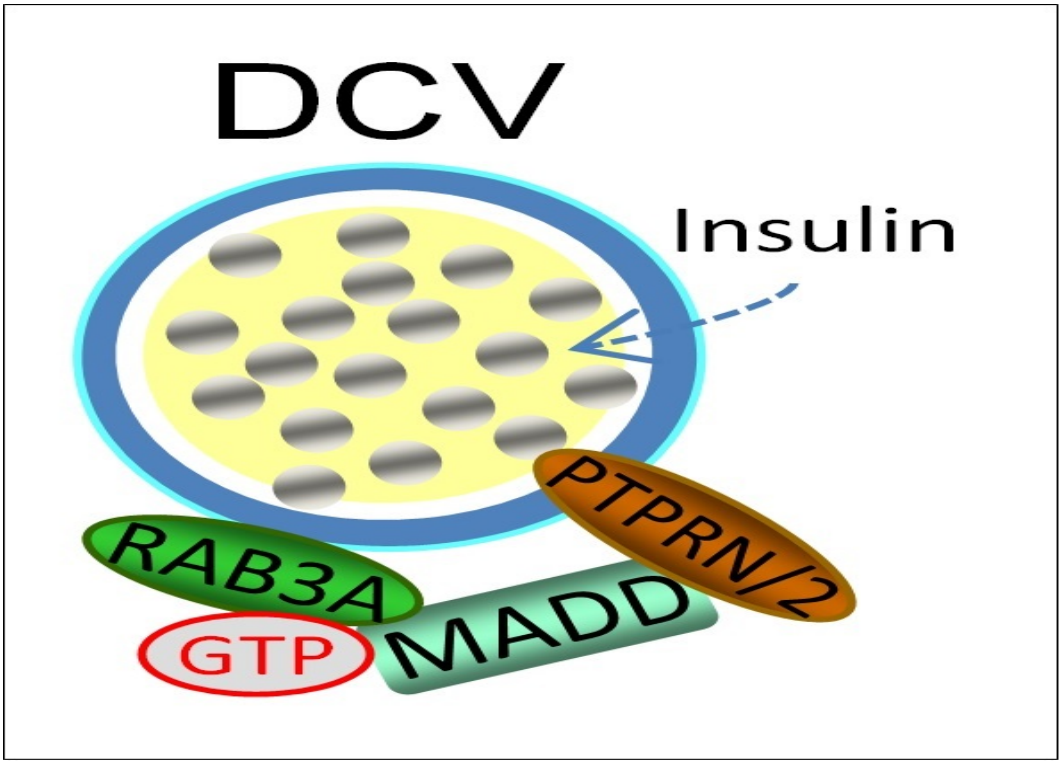

Figure 2: PTPRN and PTPRN2 are dense core vesicle (DCV) that interact with the a complex in the pancreatic islet cells that includes MADD and RAB3A/3C. This complex influences the exocytosis of insulin-containing DCV.

Table 4. SNP-SNP Pairwise Gene-Gene Interactions in AADM Study using MDR method

Gene 1 (SNPs)	Gene 2 (SNPs)	Min. of p values	Significant level
PTPRN (rs2335845)	*MADD* (rs326217)	2.28×10^{-6}	2.42×10^{-6}
PTPRN (rs10245268)	*RAB3C* (rs4700305)	1.04×10^{-10}	1.72×10^{-7}
RAB3C (rs292968)	*MADD* (rs41301449)	6.57×10^{-8}	1.58×10^{-5}

5. Discussion.

Comparing with existing genetic functional methods, such as in Lee and Park (2014), and Bochdanovits et al. (2008) and references there in, the statistic is asymptotically either a chi-squred distribution for small number of genes, or a chi-squred mixture distribution for large number of genes, in the latter case the cut-off point for testing the null hypothesis is not easy to compute, often relay on simulation methods, which is not convenient to use, we propose using Kullback-Leibler divergence method, an asymptotic postive normal distribution, overcomes these disadvantages, and make it more practical to use. We encourage to use in first screen for gene-gene interactions in genome wide level.

Gene - gene interactions have been proposed to underlie some of the missing heritability in GWAS for complex disease. Identifying gene-gene interactions in genome-wide level remains a great challenge because of computational difficulty, type 1 errors, multiple testing burden, and study power. Using genotype data to focus on a few specific genomic regions, such as gene pathway or previous studies, to detect gene-gene interactions is much more feasible, as in the real data analysis implemented. Based on earlier work, it was known that the Rab3a, MADD, and Ptprn/2 proteins form a compound that affects insulin release. Grounding

a test of gene-gene dependences in a functionally-derived hypothesis is a reasonable approach. The epistasis of two or more SNP-SNP interactions within genes is often used to instead detecting gene-gene interaction, which is described as the deviation from additivity in a linear/logistic statistical model, but this method is more likely to identify genomic hotspots for human disease, and is not likely to sufficiently capture gene-gene interactions (Hartwig 2013). Here, we propose an asymptotic positive normal statistical method to detect the degree of SNP-SNP dependence in gene-based analysis. We implemented this method in two simulation studies and in the AADM study. Comparing this method with others, ours has a reasonable computation time (the total user: 20443.25; system CPU: 4.43; and real elapsed time: 20467.57 for 2000 samples and 1500 SNPs in a windows-based R) and memory used 39.4mb. Given a sample size of 1000 and the degree of dependence of ~ 0.05, the type 1 error remains $< 5\%$ with power > 0.80 in simulation studies. Estimated values are sensitive to the total of number of included SNPs. From Figure 1, when the total number of SNPs is more than 100, the estimated D_0 and λ values are near stable. We suggest that if study sample size is near 1000, the minimum total number of SNPs should be at least 100. We recognize a few limitations of this method. First, this method is developed only for two genes. A method to detect dependences between more than two genes is being developed. Second, the method requires that the SNPs within a gene are not in linkage disequilibrium, requiring a pruning step prior to implementation. Finally, the computation time could still be improved.

In conclusion, we developed a new method to detect gene-gene dependences at the gene level by using asymptotic positive normal distribution (Kullback-Leibler

statistic). In a simulation study under the condition of no dependences, type 1 error rates were $< 5\%$. In a simulation study in the presence of dependences, the power was > 0.80. In an implementation of this method in real data, we evaluated genes (*RAB3A/3C, MADD, PTPRN/2*) encoding proteins that form a complex that affect release of insulin. We identified significant pair-wise gene-gene dependences among these genes. These gene-gene interactions were also identified using a SNP-based MDR method. For accurate estimation, a minumum 100 SNPs within two genes is required when study sample size is $\sim$1000.

Acknowledgement: This work is supported in part by the National Center for Research Resources at NIH grant 2G12RR003048, and by the Center for Research on Genomics and Global Health (CRGGH) at NHGRI/NIH.

Reference

Anno, S., Abe, T., Yamamoto, T. (2008). Interactions between SNP alleles at multiple loci contribute to skin color differences between caucasoid and mongoloid subjects. *International Journal of Biological Sciences*, **4**(2), 81-86.

Aschard, H., Gusev, A., Brown, R., Pasaniuc, B. (2015). Leveraging local ancestry to detect gene-gene interactions in genome-wide data. *BMC Genetics*, **16**:124.

Bochdanovits, Z., Sondervan, D., Perillous, S., van Beijsterveldt, T., Boomsma, D., Heutink, P. (2008). Genomewide prediction of functional gene-gene interactions inferred from patterns of genetic differentiation in mice and men. *PLoS ONE.*, **3**:e1593.

Buil, A., Brown, A.A., Lappalainen, T., Viñuela, A., Davis, M.N., Zheng, H-F., Richards, J.B., Glass, D., Small, K.S., Durbin, R., Spector, T.D., Dermitzakis, E.T. (2015). Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nature Genetics*, **47**, 88-91.

Caromile L.A., Oganesian A., Coats S.A., et et (2010). the Neurosecretory Vesicle protein phogrin functions as a phosphatidylinositol phosphatase to regulate insulin secretion. *J. of Biological Chemistry*,**285**:10487-10496

Chen G., Ramos E., Adeyemo A., Shriner D., Zhou J., Doumatey A., Huang H., Erdos M., Gerry N., Herbert A., Bentley A., Xu H., Charles B., Christman M.. Rotimi N. (2011). UGT1A1 is a Major Locus Influencing Bilirubin Levels in African Americans, *European Journal of Human Genetics*, **20**(4): 463-468

Chung, Y., Lee, S.Y., Elston, R.C., Park, T. (2007). Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. *Bioinformatics*. **23**, 7176.

Coppola T., Perret-Menoud V., Wang J., et al. The death domain of Rab3 guanine nucleotide exchange protein in GDP/GTP exchange activity in living cells. *Biochem J.*, **362**, 273-239

Cordell, H.J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**(6), 392-404.

Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., Shi, T., Huang, W., Li, Y. (2008). Exploration of gene-gene interaction effects using entropy-based methods. *Eur J Hum Genet.*, **16**, 229235.

Engelgau, M.M., Narayan, K.M., Kerman, W.H. (2000) Screening for type 2 diabetes. *Diabetes Care*,**23**(10), 1563-1580

Fan, R.Z., Wang, Y.F., Boehnke, M., Chen, W., Li, Y., Ren, H.B., Lobach, I., and Xiong, M.M. (2015) Gene level meta-analysis of quantitative traits by functional linear models. *Genetics* 200 (4):1089-1104.

Ferreira, T., Donnelly, P., Marchini, J. (2007). Powerful Bayesian gene-gene interaction analysis. *Am J Hum Genet.* S81:32.

Gilbert-Diamond, D., Moore, J.H. (2011). Analysis of gene-gene interactions, *Curr Protoc Hum Genet.*, Unit1.14. doi:10.1002/0471142905.hg0114s70.

Hartwig P.F. (2013) SNP-SNP interactions: Focusing on Variable Cording for Complex Models of Epistasis. *Genet Syndr Gene Ther*,**4**9, 189

Kasai K., Ohare-Imaizumi M., Takahashi N., et al. (2005) Rab27a mediates the tight docking of insulin granules onto the plasma membrane during glucose stimulation. *J Clin Invest*, **115**, 388-396

Kim, Y., Park, T. (2015). Robust gene-gene interaction analysis in genome wide association studies. *PLOS One*, 10(8): e0135016.

Koo, C.L., Liew, M.J., Mohamad, M.S., Salleh, A.H.M. (2013). A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *BioMed Research International*, Article ID 432375.

Kooperberg, C., Ruczinski, I. (2005). Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol.*, **28**,157170.

Lee, S.Y., Chung, Y., Elston, R.C., Kim, Y., Park, T. (2007). Log-linear model based multifactor-dimensionality reduction method to detect gene-gene interactions. *Bioinformatics*, **23**, 25892595.

Lee, S., Kwon, M-S., Oh, J.M., Park, T. (2012). Genegene interaction analysis for the survival phenotype based on the Cox model. *Biometrics*, **28**(18), 582-538.

Lee, J-H and Park, C-S. (2014). Gene - Gene interactions among MCP genes polymorphisms in asthma. *Allergy Asthma Immunol Res.*, **6**(4), 333340.

Li, J., Tang, R., Biernacka, J.M., Andrade M.d. (2009). Identification of gene-gene interaction using principal components. *BMC Proceedings*, **3**(Suupl 7):S78.

Li, J., Zhong, W., Li, R., Wu, R. (2014). A fast algorithm for detecting gene-gene interactions in geneome-wide association studies. *Annals of Applied*

Statistics, **8**(4), 2292-2318.

Lin, H-Y., Amankwah, E.K., Tseng, T-S., Qu, X., Chen, D-T., Park, J.Y. (2013). SNP-SNP interaction network in angiogenesis genes associated with prostate cancer aggressiveness.*PLoS ONE*, **8**(4): e59688. doi:10.1371/journal.pone.0059688

Lou, X.Y., Chen, G.B., Yan, L., Ma, J.Z., Mangold, J.E., Zhu, J., Elston, R.C., Li, M.D. (2008). A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. *Am J Hum Genet.*, **83**, 457467.

Martin, E.R., Ritchie, M.D., Hahn, L., Kang, S., Moore, J.H. (2006). A novel method to identify gene-gene effects in nuclear families: the MDR-PDT. *Genet Epidemiol.*, **30**,111123.

Moore, J.H. (2004). Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert Rev Mol Diagn.* **4**,795803.

Morris A.P.,, Voight B.F., Teslovich T.M., et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet*, **44**(9):981-990

Motsinger, A., Lee, S., Mellick, G., Ritchie, M. (2006). GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. *BMC Bioinformatics.*, **7**, 39.

Onay, V.$\ddot{U}$., Briollais, L., Knight, J.A., Shi, E., Wang, Y., Wells, S., Li, H., Rajendram, I., Andrulis, I.L., Ozcelik, H. (2006). SNP-SNP interactions in breast cancer susceptibility. *BMC Cancer*, **6**: 114.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, **15**, 1191-1253.

Park, M.Y., Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*. **9**, 3050.

Purcell, S., Neale, B., Todd-Brown, K., et al (2007). PLINK: a toolset for whole-genome associaiton and population-based linkage analysis. *American J. of Human Genetics*, 4 10

Rotimi, C.N., Dunston, G.M., Berg, K. et al. (2001). In search of susceptibility genes for type 2 diabetes in West Africa: the design and results of the first phase of the AADM study.]it Ann Epidemiol, **11**, 51-58.

Serfling, R. (1982). *Approximation Theorems of Mathematical Statistics*, Wiley.

Tsai, C.T., Hwang, J.J., Ritchie, M.D., Moore, J.H., Chiang, F.T., Lai, L.P., Hsu, K.L., Tseng, C.D., Lin, J.L., Tseng, Y.Z. (2007). Renin-angiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: detection of high order gene-gene interaction. *Atherosclerosis.*, *195*, 172180.

Van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*, Springer.

Xu, K., Jin, L., Xiong, M. (2017). Functional regression method for whole genome eQTL epistasis analysis with sequencing data. *BMC Genetics*, 18:385 DOI 10.1186/s12864-017-3777-4.

Yang, P., Ho, J.W.K., Yang, Y.H., Zhou, B.B. (2011). Gene-gene interaction filtering with ensemble of filters. *BMC Bioinformatics*, **12** (Suppl 1): S10.

Yuan A, Cheng G, Zhou Y, Rotimi C. (2012). Combined Rare and Common Variants Association Analysis: A Noval Approach to Evaluate Genetic Variants,

In revision to submission.

Zhao, J., Zhu, Y., Xiong, M. (2016). Genome-wide gene-gene interaction analysis for next-generation sequencing. *European Journal of Human Genetics*, **24**, 421-428.

Zhu X, Feng T, Li Y, Lu Q, Elston RC. (2010). Detecting rare variants for complex traits using family and unrelated data. *Genetic Epidemiology*, **34**: 171-187.