이학석사학위논문

Identification of genes related to cancer through targeted sequencing data analysis

표적 시퀀싱 자료를 이용한 암 관련 유전자 발굴

2017 년 2 월

서울대학교 대학원

자연과학대학 통계학과

이 주 원

Identification of genes related to cancer through targeted sequencing data analysis

by

JOOWON LEE

A thesis
submitted in fulfillment of the requirement
for the degree of Master
in
Statistics

Department of Statistics
College of Natural Sciences
Seoul National University
Feb, 2017

Identification of genes related to cancer through targeted sequencing data analysis

지도교수 박태성

이 논문을 이학석사 학위논문으로 제출함

2017 년 2 월

서울대학교 대학원
자연과학대학 통계학과
이주원

이주원의 이학석사 학위논문을 인준함

2017 년 2 월

위원장	조신섭	**(인)**
부위원장	박태성	**(인)**
위원	이영조	**(인)**

Abstract

Identification of genes related to cancer through targeted sequencing data analysis

JOOWON LEE
Department of Statistics
The Graduate School
Seoul National University

Recent statistical methods for next generation sequencing (NGS) data have been successfully applied to identifying rare genetic variants associated with certain diseases. Note that most commonly used methods such as burden tests and variance-component tests rely on large sample size. However, due to a high cost of sequencing, small sample size sequencing data are popularly generated. Most existing methods are not appropriate to handle sequencing data with small samples.

In this work, we propose a new exact association test for sequencing data which does not require a large sample approximation. Our method is based upon the generalized Cochran-Mantel-Haenszel

(CMH) statistic. We applied our method to NGS data from Intraductal papillary mucinous neoplasm (IPMN) patients. These IPMN patients have the unique pancreatic neoplasm which could turn into an invasive and hard-to-treat pancreatic cancer. Through this application, we successfully identified susceptible genes that are associated with the progression of IPMN to pancreatic cancer.

Key words: NGS data analysis, Small sample association study, Fisher’s exact test, CMH statistic, IPMN

Student number: 2015-20309

Contents

List of Figures

List of Tables

1. Introduction

Many genetic studies such as genome-wide association studies (GWAS) have successfully identified genetic variants associated with complex human traits and diseases [1]. However, GWAS mainly focus on common variants which have minor allele frequencies (MAF) greater than 0.05. Thus, it is found that significant GWAS variants explain a small proportion of disease heritability, which is called the missing heritability. One of the possible reasons to explain missing heritability is about rare variants [2, 3]. Fortunately, the development of sequencing technologies has put large-scale investigation of rare variation within reach, which could contribute substantially to missing heritability [4]. It is found that rare genetic variants, defined as MAF between 0.01 and 0.05, play an important role in complex diseases [3].

Up to date, various statistical methods and strategies to test associations for rare genetic variants have been developed. Burden tests, which are one of earlier tests for rare variants, aggregate information of all rare variants in a region into a single summary variable [5, 6]. Different types of burden tests have been proposed using various genetic scores of the rare variants. For example, the cohort allelic sum test (CAST) collapses genotypes across all variants in such a way that an individual is coded as 1 if a rare allele is present at any of the variant sites and as 0 otherwise [5]. However, this approach may not fully reflect the effect emerging from the complex ensemble of multiple rare variants, because it only uses the information of presence of rare variants in the region.

The combined multivariate and collapsing (CMC) method divides rare variants into multiple classes based on their MAFs, collapses each group using CAST approach, and then uses multivariate tests such as Hotelling's T-test [6]. However, these burden tests are powerful only if most rare variants are causal and have effects in the same direction. In other words, existence of variants whose effects are in different directions can reduce power substantially. To overcome this limitation, several variance-component (VC) test based on a regression models were proposed. The Sequence Kernel Association Test (SKAT), a most widely used score-based VC test, has been shown to successfully detect multiple directional contribution from different classes of single nucleotide polymorphism (SNP)s [7].

Note that burden tests and VC tests for rare variants are based on the asymptotic tests assuming that the sample size is large enough. Due to high sequencing cost, however, small sample size sequencing data are commonly generated. These existing methods are not appropriate to handle sequencing data with small samples. To handle the next generation sequencing (NGS) data with small samples, the SKAT method needs to be modified by renormalizing moments of test statistics [8].

In this study, we propose a new approach which does not rely on the asymptotic distribution for the NGS data with small samples. We call this new method the Exact Association Test (EAT). EAT is conceptually based upon the Fisher's exact test which is commonly used for independence test in a 2×2 contingency table with small samples. A key underlying assumption of Fisher's exact test is that the four marginal sums are fixed. Under this assumption, the first cell frequency follows a hypergeometric distribution under the null hypothesis of independence. The Cochran-Mantel-Haenszel (CMH) statistic was developed to extend Fisher's exact test for a stratified 2×2 contingency tables for testing the conditional independence between two categorical variables conditioned on the third categorical variable [9, 10]. The generalized Cochran-Mantel-Haenszel (GCMH) statistic is an extension of CMH for stratified J×K contingency tables [10].

For a specific gene, the NGS data can be represented by a sequence of contingency tables. The strata variable corresponds to the subject, the row variable does to single nucleotide variant (SNV) and the column does the genotypes which represent the number of minor allele (0, 1, or 2). For example, suppose that a gene contains t SNVs. Then, the NGS data from n individuals can be summarized into the $n \times t \times 3$ contingency tables. The GCMH statistic can be applied to these stratified contingence tables. Note that this GCMH statistic for testing independence between SNVs and the number of minor allele. That is, it tests whether t SNVs have similar distributions in terms of MAFs. However, this GCMH does not provide any information about the association between the gene and the disease status, say case and control. Thus, we propose deriving the GCMH statistics separately from case and control groups and using the difference or ratio as a test statistic. If these two GCMH statistics differ greatly between case and control groups, then the gene is expected to be strongly associated with the disease status.

Section 2 provides a detailed description on EAT statistic and summarizes how to compute the p-values for the significance testing. We then apply our EAT to the analysis of targeted sequencing data from Intraductal papillary mucinous neoplasm (IPMN) patients. These IPMN patients have unique pancreatic neoplasm which could turn into an invasive and hard-to-treat pancreatic cancer [11]. Through this application, we demonstrate that our proposed EAT method can successfully identify susceptible genes that are associated with the progression of IPMN to pancreatic cancer.

2. Materials and Methods

2.1 Materials

Our targeted sequencing data were generated using the Illumina NextSeq500 platform. Surgical paraffin-embedded IPMN samples of 44 subjects were obtained from Seoul National University hospital. These subjects consist of 21 cases of high grade (just before developing pancreatic cancer) and 23 controls of low grade (benign tumor).

The demographic and clinical characteristics of the 44 subjects are shown in *Table 1*. Categorical variables were compared using the χ^2 test or Fisher's exact test between case and control groups. Continuous variables were compared using Student's t-test or Wilcoxon's rank sum test. Except *Mural Nodule* and *Invasiveness*, there are no significant differences between case and control groups. *Mural Nodule* is known as a potential predictor of malignant neoplasm [12], and *Invasiveness* presents a grade status (0=no, 1=yes) just before pancreatic cancer.

Table 1. Demographic and clinical characteristics of study patients at baseline

	Total (n=44)	**Case (n=21)**	**Control (n=23)**	$\boldsymbol{P}$ **– value**
Age	64.57 (8.3)	64 (9.4)	65.09 (7.2)	0.668
Sex (Male : Female)	28:16	15:6	13:10	0.476
Invasiveness ratio (Invasive : Noninvasive)	10:34	9:12	1:22	0.003
Mural Nodule (Yes : No)	19:24	13:8	6:16	0.032
Recurrence (Yes : No)	34:9	4:17	0:22	0.044
Survival (Yes : No)	34:9	17:4	17:5	1
CEA	2.46 (2.6)	2.90 (3.4)	2.04 (1.4)	0.293
CA19-9	60.61 (280.9)	2.89 (400.2)	2.04 (11.0)	0.061

Each patient has the targeted sequencing data of 411 genes which are known to be related to cancer in general, not necessarily to pancreatic cancer. The total number of SNV is 8325. The number of SNVs in a gene ranges from 1 to 188, and the median is 15.

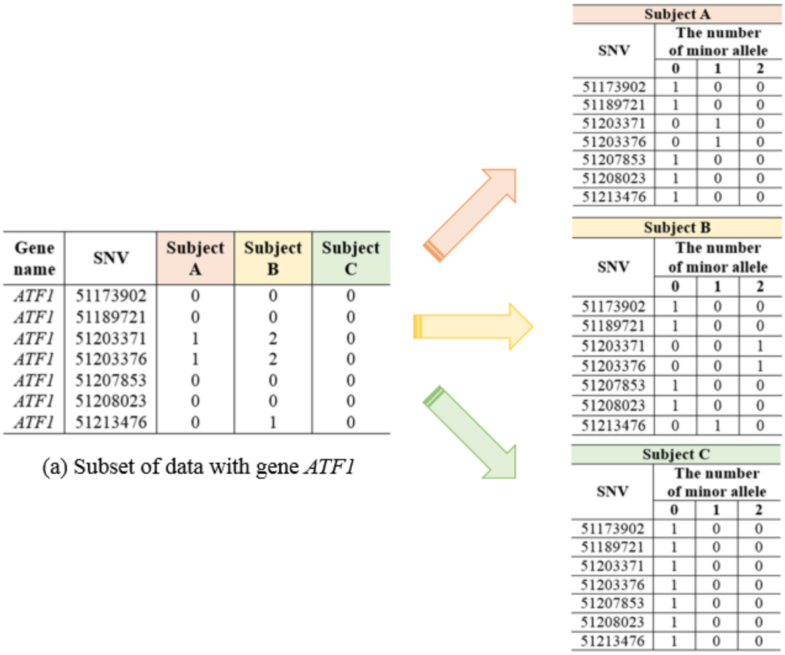
2.2 Methods

We construct a stratified categorical data as follows. For a given gene with *t* SNVs, define a *t×3* contingency table for each subject, where the rows and columns represent SNV of the gene and the number of minor allele, respectively. More precisely, for subject *i* and a specific gene with *t* SNVs, the corresponding *t×3* contingency table can be constructed, as shown in *Table 2*. Note that the cell count $\mathrm{n_{ijk}}$ has a value 1 if subject *i* has a minor allele count *k* at SNV *j* for $\mathrm{i} = 1, \cdots \mathrm{n}, \mathrm{j} = 1, \cdots, \mathrm{t}, \mathrm{k} = 0, 1, 2$.

Table 2. A strata representing subject *i* for a specific gene with *t* SNVs

SNV	The number of minor allele			Total
	0	1	2	
1	n_{i10}	n_{i11}	n_{i12}	*1*
⋮	⋮	⋮	⋮	*1*
t	n_{it0}	n_{it1}	n_{it2}	*1*
Total	$n_{i.0}$	$n_{i.1}$	$n_{i.2}$	t

For example, consider a gene *ATF1* containing seven SNVs. *Figure 1 (a)* shows the number of minor allele for three subjects A, B, and C. From these data, three *7×3* contingency tables can be constructed as shown in *Figure 1 (b)*.



Gene name	SNV	Subject A	Subject B	Subject C
ATF1	51173902	0	0	0
ATF1	51189721	0	0	0
ATF1	51203371	1	2	0
ATF1	51203376	1	2	0
ATF1	51207853	0	0	0
ATF1	51208023	0	0	0
ATF1	51213476	0	1	0

(a) Subset of data with gene *ATF1*

Subject A			
SNV	The number of minor allele		
	0	1	2
51173902	1	0	0
51189721	1	0	0
51203371	0	1	0
51203376	0	1	0
51207853	1	0	0
51208023	1	0	0
51213476	1	0	0

Subject B			
SNV	The number of minor allele		
	0	1	2
51173902	1	0	0
51189721	1	0	0
51203371	0	0	1
51203376	0	0	1
51207853	1	0	0
51208023	1	0	0
51213476	0	1	0

Subject C			
SNV	The number of minor allele		
	0	1	2
51173902	1	0	0
51189721	1	0	0
51203371	1	0	0
51203376	1	0	0
51207853	1	0	0
51208023	1	0	0
51213476	1	0	0

(b) The format represented on the contingency table

Figure 1. Transformation of a subset of data (a) into contingency tables for each subject

Let $\mathbf{n_i} = (n_{i10}, n_{i11}, \cdots, n_{it2})'$ denote the $(3t) \times 1$ vector of observed frequencies, $\mathbf{n_{ij.}} = (n_{i1.}, n_{i2.}, \cdots, n_{it.})'$ denote the vector of the row marginal total number, $\mathbf{n_{i.k}} = (n_{i.0}, n_{i.1}, n_{i.2})'$ denote the vector of the column marginal totals, and $\mathrm{n_{i..}}$ denote the overall marginal total. Note that all row marginal totals $\{\mathbf{n_{ij.}}\}$ are 1, and $\mathrm{n_{i..}}$

has the value *t*, the number of SNVs. Under the null hypothesis of independence between SNV and the number of minor allele, $\mathbf{n_i}$ in the *Table 2* follows a multiple hypergeometric distribution when the marginal totals are fixed, as in Fisher's exact test.

$$\mathbf{n_i} = (n_{i10}, n_{i11}, \cdots, n_{it2})' \sim \mathrm{H}(t, \{\mathbf{n_{ij.}}\}, \{\mathbf{n_{i.k}}\})$$

Then, the GCMH statistic can be derived from $\mathbf{n_i}$ as follows, where $\otimes$ denotes the Kronecker product, i.e, defined for matrices $\mathbf{A} = \{\mathrm{a_{ij}}\}$ and $\mathbf{B}$ of any order to be $\mathbf{A} \otimes \mathbf{B} = \{\mathrm{a_{ij}}\mathbf{B}\}$ [13], and the linear transformation matrix $\mathbf{A}$ eliminates the last row and last column of the contingency table.

$$\mathrm{GCMH} = \frac{\left(\mathbf{A}\sum_\mathrm{i} \mathrm{n_i} - \mathbf{A}\mathrm{E}(\sum_\mathrm{i} \mathrm{n_i})\right)'\left(\mathbf{A}\sum_\mathrm{i} \mathrm{n_i} - \mathbf{A}\mathrm{E}(\sum_\mathrm{i} \mathrm{n_i})\right)}{\mathbf{A}\mathrm{Var}(\sum_\mathrm{i} \mathrm{n_i})\mathbf{A}'},$$

$$\mathbf{A}: (\mathrm{I_{t-1}}, 0_{\mathrm{t-1}}) \otimes (\mathrm{I_2}, 0_2)$$

A large value of the GCMH statistic indicates an association between the number of minor allele and the SNV in a specific gene in at least one of the subjects.

In genetic association studies, we need to identify the genes associated with a certain phenotype of interest such as disease status, say, case and control. Our GCMH statistics does not use any disease status information. In order to take the phenotype information, we propose computing the GCMH statistics from case and control group separately, and considering the ratio. We denote the two GCMH as $\mathrm{CMH}_{\mathrm{case}}$ and $\mathrm{CMH}_{\mathrm{control}}$ and our proposed gene-based Exact Association Test (EAT) T as

$$\mathrm{T} = \log \mathrm{CMH}_{\mathrm{case}} - \log \mathrm{CMH}_{\mathrm{control}}$$

This test statistic needs to be computed for each gene X. The large absolute value of T statistic implies that the strength of dependence between t SNVs and the number of minor allele differs between case and control groups. Thus, the gene with large absolute value of T statistic is expected to be strongly associated with the disease status. We obtain p-value by a permutation procedure. Genes that have smaller p-value than the pre-specified significance level can be identified as associated to a disease status, e.g., the progression of IPMN to pancreatic cancer in our study.

3. Results

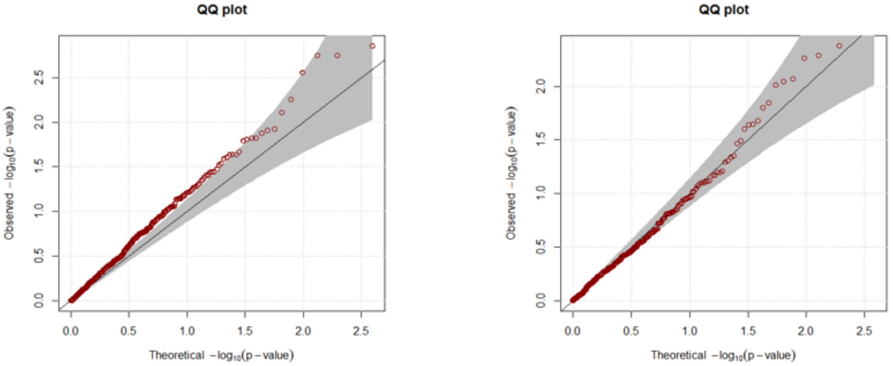
We applied the proposed EAT to 395 genes by excluding the genes which cannot compute the ratio values. If any gene has only 1 SNV, then we cannot construct the contingency table used in EAT. In this case, we simply examine the significance of the association between the disease status and the number of minor allele through the Fisher's exact test.

Through 10,000 times permutation, our EAT identified 31 significant genes at a significance level of 0.05. The part of the results are represented in *Table 3*. *Table 3* shows the top 5 p-values from EAT, SKAT, and SKAT-O methods. It also provides p-values of four known oncogenes related to pancreatic cancer. It was recently reported that mutations in *KRAS* are strongly related to cancer development and progression [14], and only EAT method could find *KRAS* as a significant gene. However, we have lots of genes so that no gene could detect through multiple comparisons methods.

Table 3. P-values from EAT and competed methods with top 5 genes and oncogenes

Gene name	EAT			SKAT			SKAT-O		
	P-value	Bonferroni	FDR	P-value	Bonferroni	FDR	P-value	Bonferroni	FDR
FLT1	0.0014	0.5530	0.2370	0.1484	1.0000	0.6679	0.1187	1.0000	0.4581
AURKB	0.0018	0.7110	0.2370	0.0591	1.0000	0.6679	0.0008	0.3493	0.2513
KMT2D	0.0018	0.7110	0.2370	0.0818	1.0000	0.6679	0.0137	1.0000	0.3179
ITGA9	0.0028	1.0000	0.2765	0.1337	1.0000	0.6679	0.1854	1.0000	0.5044
CYP2C19	0.0056	1.0000	0.4424	0.0984	1.0000	0.6679	0.0261	1.0000	0.3247
KRAS	0.0251	1.0000	0.5405	0.7204	1.0000	0.7751	0.3209	1.0000	0.5709
TP53	0.1986	1.0000	0.7173	0.1742	1.0000	0.6679	0.2309	1.0000	0.5231
GNAS	0.5972	1.0000	0.9170	0.4262	1.0000	0.6691	0.5970	1.0000	0.7309
CDH1	0.9630	1.0000	1.0000	0.6989	1.0000	0.7751	0.7775	1.0000	0.8434

A QQ plot of our EAT is shown in *Figure 2 (a)* showing an inflation pattern. Since our NGS data is targeted sequencing data, it contains many genes known or suspected to be associated with the cancer. In order to investigate whether the inflation is caused by causal association or by false positive, we permuted the disease status (case and control) from our data and then generated QQ plots. All QQ plots showed a similar pattern without any inflation. *Figure 2 (b)* shows one of QQ plots. Since there was no inflation after permutation, the inflation pattern in *Figure 2 (a)* was indeed resulted from the causal genes related to cancer.

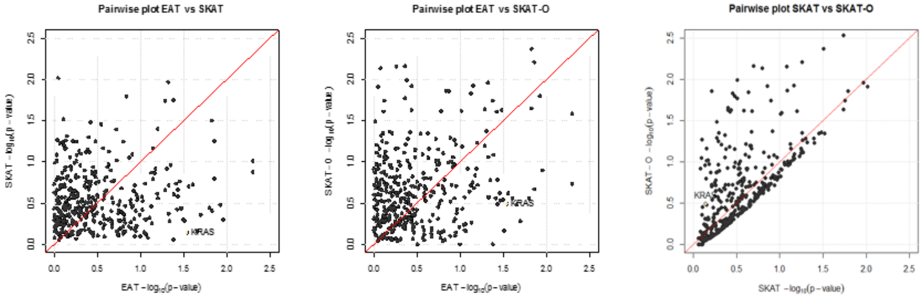


(a) A QQ plot from original data

(b) A QQ plot from the data permuted disease status

Figure 2. QQ plot

Pairwise scatter plots of EAT with SKAT and SKAT-O are shown below in *Figure 3 (a)* and *(b)*, and we do not observe any clear patterns. *Figure 3 (c)* shows the pairwise scatter plots of SKAT with SKAT-O. The p-values of some genes are similar, but in other cases, SKAT-O is more powerful than SKAT. That’s because SKAT-O combines the burden test and the nonburden SKAT. SKAT-O behaved like SKAT when SKAT was more powerful than the burden tests, and they behaved like burden tests when the burden tests were more powerful than SKAT [8].



(a) Pairwise scatter plot of EAT vs SKAT

(b) Pairwise scatter plot of EAT vs SKAT-O

(c) Pairwise scatter plot of SKAT vs SKAT-O

Figure 3. Pairwise scatter p-value plots of EAT and other competed methods

4. Discussion

We have proposed a new test EAT for association and applied to IPMN data. EAT is an exact association test which does not require a large sample approximation. Our method is conceptually based upon the Fisher's exact test, and computed our test statistics using the GCMH.

As shown in *Figure 3*, EAT provides different p-values from those of SKAT and SKAT-O, which is mainly because EAT and SKAT (or SKAT-O) use different test statistics to detect significant genes from NGS data. Our proposed EAT uses the GCMH statistics for an array of contingency table between the number of minor allele and SNV. Under the assumption of randomness within each group, EAT is derived under the hypergeometric distributional assumption conditioned on marginal totals. Thus, we can compute the partial average association from each group. The ratio of two GCMHs from case and control groups is then used to compare the extent of partial association between case and control groups, and the p-value is obtained by a permutation test. On the other hand, SKAT method is based on the regression model and uses a variance-component test. SKAT evaluates the significance of genes to rely on the regression coefficients of variants using score test statistics. Score test statistics follow the asymptotic chi-square distribution under the null hypothesis.

This theoretical difference suggests that in some cases our method could be more suited than the other two methods. EAT is valid for all sample sizes, so can be more accurate than SKAT in the small sample study since SKAT relies on an approximation that becomes exact as the sample size grows to infinity. Indeed, as indicated in *Figure 3*, among the three methods, only EAT successfully identified the gene called *KRAS* which is well known to be related to pancreatic cancer. This illustrates that our newly proposed method could effectively identify susceptible genes that are associated with the progression of IPMN to pancreatic cancer.

Despite the distinctive performance of EAT in distinguishing groups of different distributions, it has the following limitations to be improved in a future work: (1) It provides hypothesis testing procedures only; (2) EAT may be insensitive when associations vary in direction across subjects within a group; (3) We could not develop analytic derivation of distribution of our EAT statistic under null hypothesis.

Lastly, we remark on directions of future studies. First, we will compare the performance of EAT with other existing tests for the NGS data with small samples through simulations. Second, we can incorporate other types of GCMH statistics such as mean score or correlation CMH in our framework. The resulting test statistics may reflect further biological information so that they could improve EAT in terms of power. Lastly, we will also apply our methods to the study of other NGS data in our future research.

Bibliography

1. M. C. Wu, P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter, et al., "Powerful SNP-set analysis for case-control genome-wide association studies," Am J Hum Genet, vol. 86, pp. 929-42, Jun 11 2010.

2. G. Gibson, "Hints of hidden heritability in GWAS," Nat Genet, vol. 42, pp. 558-60, Jul 2010.

3. T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, et al., "Finding the missing heritability of complex diseases," Nature, vol. 461, pp. 747-53, Oct 8 2009.

4. D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, "The next-generation sequencing revolution and its impact on genomics," Cell, vol. 155, pp. 27-38, Sep 26 2013.

5. S. Morgenthaler and W. G. Thilly, "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST)," Mutat Res, vol. 615, pp. 28-56, Feb 3 2007.

6. B. Li and S. M. Leal, "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data," Am J Hum Genet, vol. 83, pp. 311-21, Sep 2008.

7. M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test," Am J Hum Genet, vol. 89, pp. 82-93, Jul 15 2011.

8. S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, et al., "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies," Am J Hum Genet, vol. 91, pp. 224-37, Aug 10 2012.

9. C. S. Davis, Statistical methods for the analysis of repeated measurements: Springer Science & Business Media, 2002.

10. J. R. Landis, E. R. Heyman, and G. G. Koch, "Average partial association in three-way contingency tables: a review and discussion of alternative tests," International Statistical Review/Revue Internationale de Statistique, pp. 237-254, 1978.

11. R. Salvia, C. Fernandez-del Castillo, C. Bassi, S. P. Thayer, M. Falconi, W. Mantovani, et al., "Main-duct intraductal papillary mucinous neoplasms of the pancreas: clinical predictors of malignancy and long-term survival following resection," Ann Surg, vol. 239, pp. 678-85; discussion 685-7, May 2004.

12. J. Y. Jang, T. Park, S. Lee, M. J. Kang, S. Y. Lee, K. B. Lee, et al., "Validation of international consensus guidelines for the resection of branch duct-type intraductal papillary mucinous neoplasms," Br J Surg, vol. 101, pp. 686-92, May 2014.

13. H. V. Henderson, F. Pukelsheim, and S. R. Searle, "On the history of the Kronecker product," Linear and Multilinear Algebra, vol. 14, pp. 113-120, 1983.

14. F. Schonleben, J. D. Allendorf, W. Qiu, X. Li, D. J. Ho, N. T. Ciau, et al., "Mutational analyses of multiple oncogenic pathways in intraductal papillary mucinous neoplasms of the pancreas," Pancreas, vol. 36, pp. 168-72, Mar 2008.

초 록

차세대 시퀀싱 (NGS) 데이터에 대한 최신 통계적 분석 방법들은 특정 질병과 관련된 희귀 유전변이를 성공적으로 규명하였다. 가장 보편적으로 사용되는 burden test 와 variance-component test 는 대규모 표본 데이터에 의존하는 방법이다. 그러나 시퀀싱 기술의 높은 비용으로 인해 소규모 표본의 시퀀싱 데이터가 주로 생산되므로, 소규모 표본의 시퀀싱 데이터에 적용 가능한 통계적 분석 방법이 필요하다.

본 연구에서 우리는 대표본 근사에 의존하지 않는 새로운 정확한 연관성 검정 방법 (exact association test)을 제안한다. 이 방법은 일반화된 Cochran-Mantel-Haenszel (CMH) 통계량에 근거하고 있다.

우리는 본 방법을 유두상 점액종양 (Intraductal papillary mucinous neoplasm; IPMN) 환자들에게서 얻은 NGS 데이터에 적용하였다. 유두상 점액종양은 췌장에 발생하는 양성 종양으로, 일부 환자들에게서 치료가 어렵고 생존율이 극히 낮은 췌장암으로 진행된다. 데이터에의 적용을 통해 우리는 유두상 점액종양에서 췌장암으로의 진행과 관련된 유전자를 효과적으로 규명해내었다. 이는 췌장암에 걸릴 가능성이 높은 환자들을 사전에 분류하고, 외과적 수술요법 등을 통해 위험요인을 미리 제거함으로써 생존율을 높이는 데 기여할 수 있을 것이다.

주요어: 차세대 시퀀싱 데이터 분석, 소규모 표본 연관성 분석, Fisher 의 정확한 검정, CMH 통계량, 유두상 점액종양 (IPMN)

학 번: 2015-20309