A Data-adaptive SNP-Set-based Association Test of Longitudinal Traits and the extension to Test on Gene Pathway

by

Yang Yang, M.S

APPROVED:

Thesis chair, PHD

Prof 2, PHD

Minor Prof, PHD

Breadth Prof, PHD

DEDICATION

Persistent support from my family members:

Nainan Hei

&

Tianpeng Yang and Qi Lu

A Data-adaptive SNP-set-based Association Test of Longitudinal Traits and the extension to Test on Gene Pathway

by

Yang Yang, M.S

Presented to the Faculty of The University of Texas
School of Public Health
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS
SCHOOL OF PULIC HEALTH
Houston, Texas
November, 2015

ACKNOWLEDGEMENTS

Great thanks to my dissertation adviser Dr. Peng Wei, as he guided me ever from 2011, put countless efforts in training me to be a countable person, and then a qualified Ph.D. I also appreciate great helps from my dissertation committee members: Dr. Han Liang, Dr. Alanna C. Morrison and Dr. Yun-Xin Fu. They are talented experts in their fields and provided me with enormous valuable advice towards my research and writings.

A Data-adaptive SNP-set-based Association Test of Longitudinal Traits and the extension to Test on Gene Pathway

Yang Yang, M.S
The University of Texas
School of Public Health, 2014

Thesis Chair, Peng Wei, PhD

Prof 2, Liang Han, PhD

Minor Prof, Alanna C. Morrison, PhD

Breadth Prof, Yun-Xin Fu, PhD

Contents

List of Tables

List of Figures

1 Background

Genome-wide association studies (GWASs) has been popular ever since 2007, and till now hundreds of GWAS have been published already [13]. The most popular approach in GWAS is to test the association between complex traits and single nucleotide variant (SNV) one by one, then select the the SNVs meeting a stringent significance level after multiple testing error correction, such as Bonferroni and FDR methods [A CITE HERE]. However, this strategy will suffer from low power when the minor allele frequency (MAF) of the SNV is low (between 1% and 5%), and as a result the signal contained within the SNV is weak [17]. Such a case becomes even more severe a problem for rare variants (RVs), which usually has MAF below 1% [2]. Although with extremely low MAF, we cannot underestimate RVs' important effects underlying disease risk, which are usually functional and deleterious; RVs also bring over larger effect size than common variants [7, 2, 17]. Therefore, developing new association test tailored to SNVs with low MAF and RVs has been very active research area in recent years. Due to the nature of low MAF, either increasing case sample size or aggregating information across multiple variants in an analysis set (e.g. gene) is expected to achieve a practically acceptable power [5, 3, 2, 17]. As increase sample size is usually expensive and demanding, e.g. more than 25,000 cases will be required, advances in gene-based and sets of functionally related genes tests are major directions people have been investigating towards [21, 15, 17]. Sets of genes can be defined by, e.g. Gene Ontology terms, protein-protein interaction, canonical gene signal pathways, gene expression networks, etc [17, 6, 19, 18].

While many GWASs have been performed in cohorts, they collected data across multiple time points for each individual [1, 10, 11, 12, 16]. However, the longitudinal information has not been fully utilized as the majority of current association tests only used either the baseline measurement or average measurement for each individual[16, 10, 11, 12]. Compared to the total number of GWASs, very few studies involved longitudinal data analysis. One such study on smoking and nicotine dependence by Belsky et.al have data from a 4-decade longitudinal study, and they used generalized estimating equation model to analyze the panel data account for correlation within subject [4]. There are also several studies on Alzheimer's Disease (or more specifically ADNI-1 data collected by Alzheimer's Disease Neuroimaging Initiative) involving the analyses of longitudinal phenotypic information collected at multiple time points [?, ?, ?]. Increased power coming from longitudinal data seems intuitive, and recently this fact has been discussed in depth by either simulation study and/or real data analysis [20, 8]. Depending on specific parameters settings in simulation studies and case by case for real data analysis, the power gain from longitudinal data analysis as compared to baseline data analysis can range from moderate to significant amount. [20, 8].

Extending the gene-based association test to sets of multiple related genes could return more biological meaningful inference, as in vivo, there are usually multiple genes working together to fulfill a biological function [A CITE HERE], analyzing "co-workers" genes together tends to identify more signal hidden from or attenuated by each single gene . Complex disease are known to have a combination of genetic factors in addition to environmental, lifestyle factors, and their interactions [A CITE HERE]. Thus by investigating into the combination of genes, more evidence could be extracted as risk altering factors contributing to a specific disease. Among association tests on sets of functional related genes, gene pathway based association test is probably the most popular one [6, 18]. There are two major categories of testing: self-contained approach and competitive approach [18, 14, 9]. The difference between the two major tests lies in the null hypothesis each test makes: self-constrained approach hypothesizes there is no gene in the gene set associated with the phenotype while competitive approach hypothesizes the same level of association of a gene set with the given phenotype as the complement of the gene set.

Meanwhile, rapid advances in rare-variant gene-based and pathway-based tests are paving the way for more powerful exome sequencing studies for complex diseases.

2 Specific Aims and Hypotheses

1. One

2. Two

3. ...

3 Data

4 Methods

4.1 Subsection one

4.1.1 Subsubsection one

4.1.2 Subsubsection two

4.2 Subsection two

4.2.1 Subsubsection one

4.2.2 Subsubsection two

5 Plan for Simulation Studies

5.1 Subsection one

5.2 Subsection two

Table 1: Example table

		A		*B*		*C*	
par	*truth*	*est*	*95% CI*	*est.*	*95% CI*	*est.*	*95% CI*
β_1	10		(,)		(,)		(,)
β_2	1		(,)		(,)		(,)
β_3	-1		(,)		(,)		(,)

5.2.1 Subsubsection one

Figure 1: Sample figure.

References

[1] Yurii S. Aulchenko, Samuli Ripatti, Ida Lindqvist, Dorret Boomsma, Iris M. Heid, Peter P. Pramstaller, Brenda W J H. Penninx, A Cecile J W. Janssens, James F.

Wilson, Tim Spector, Nicholas G. Martin, Nancy L. Pedersen, Kirsten Ohm Kyvik, Jaakko Kaprio, Albert Hofman, Nelson B. Freimer, Marjo-Riitta Jarvelin, Ulf Gyllensten, Harry Campbell, Igor Rudan, Asa Johansson, Fabio Marroni, Caroline Hayward, Veronique Vitart, Inger Jonasson, Cristian Pattaro, Alan Wright, Nick Hastie, Irene Pichler, Andrew A. Hicks, Mario Falchi, Gonneke Willemsen, Jouke-Jan Hottenga, Eco J C. de Geus, Grant W. Montgomery, John Whitfield, Patrik Magnusson, Juha Saharinen, Markus Perola, Kaisa Silander, Aaron Isaacs, Eric J G. Sijbrands, Andre G. Uitterlinden, Jacqueline C M. Witteman, Ben A. Oostra, Paul Elliott, Aimo Ruokonen, Chiara Sabatti, Christian Gieger, Thomas Meitinger, Florian Kronenberg, Angela Döring, H-Erich Wichmann, Johannes H. Smit, Mark I. McCarthy, Cornelia M. van Duijn, Leena Peltonen, and E. N. G. A. G. E Consortium . Loci influencing lipid levels and coronary heart disease risk in 16 european population cohorts. *Nat Genet*, 41(1):47–55, Jan 2009.

[2] Vikas Bansal, Ondrej Libiger, Ali Torkamani, and Nicholas J. Schork. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet*, 11(11):773–785, Nov 2010.

[3] Saonli Basu and Wei Pan. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol*, 35(7):606–619, Nov 2011.

[4] Daniel W Belsky, Terrie E Moffitt, Timothy B Baker, Andrea K Biddle, James P Evans, HonaLee Harrington, Renate Houts, Madeline Meier, Karen Sugden, Benjamin Williams, et al. Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade longitudinal study. *JAMA psychiatry*, 70(5):534–542, 2013.

[5] Marinela Capanu, Patrick Concannon, Robert W. Haile, Leslie Bernstein, Kathleen E. Malone, Charles F. Lynch, Xiaolin Liang, Sharon N. Teraoka, Anh T. Diep, Duncan C. Thomas, Jonine L. Bernstein, W. E. C. A. R. E Study Collaborative Group , and

Colin B. Begg. Assessment of rare brca1 and brca2 variants of unknown significance using hierarchical modeling. *Genet Epidemiol*, 35(5):389–397, Jul 2011.

[6] Omar De la Cruz, Xiaoquan Wen, Baoguan Ke, Minsun Song, and Dan L. Nicolae. Gene, region and pathway level analyses in whole-genome studies. *Genet Epidemiol*, 34(3):222–231, Apr 2010.

[7] Wenqing Fu, Timothy D. O'Connor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M. Leal, Stacey Gabriel, Mark J. Rieder, David Altshuler, Jay Shendure, Deborah A. Nickerson, Michael J. Bamshad, N. H. L. B. I Exome Sequencing Project , and Joshua M. Akey. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220, Jan 2013.

[8] Nicholas A Furlotte, Eleazar Eskin, and Susana Eyheramendy. Genome-wide association mapping with longitudinal data. *Genetic epidemiology*, 36(5):463–471, 2012.

[9] Jelle J. Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, Apr 2007.

[10] Iuliana Ionita-Laza, Matthew B McQueen, Nan M Laird, and Christoph Lange. Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100k scan. *The American Journal of Human Genetics*, 81(3):607–614, 2007.

[11] Yoichiro Kamatani, Koichi Matsuda, Yukinori Okada, Michiaki Kubo, Naoya Hosono, Yataro Daigo, Yusuke Nakamura, and Naoyuki Kamatani. Genome-wide association study of hematological and biochemical traits in a japanese population. *Nat Genet*, 42(3):210–215, Mar 2010.

[12] Sekar Kathiresan, Alisa K. Manning, Serkalem Demissie, Ralph B. D'Agostino, Aarti Surti, Candace Guiducci, Lauren Gianniny, Nöel P. Burtt, Olle Melander, Marju Orho-Melander, Donna K. Arnett, Gina M. Peloso, Jose M. Ordovas, and L Adrienne Cupples.

A genome-wide association study for blood lipid phenotypes in the framingham heart study. *BMC Med Genet*, 8 Suppl 1:S17, 2007.

[13] Mark I. McCarthy, Gonçalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P A. Ioannidis, and Joel N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–369, May 2008.

[14] Dougu Nam and Seon-Young Kim. Gene-set approach for expression pattern analysis. *Brief Bioinform*, 9(3):189–197, May 2008.

[15] Dalila Pinto, Alistair T. Pagnamenta, Lambertus Klei, Richard Anney, Daniele Merico, Regina Regan, Judith Conroy, Tiago R. Magalhaes, Catarina Correia, Brett S. Abrahams, Joana Almeida, Elena Bacchelli, Gary D. Bader, Anthony J. Bailey, Gillian Baird, Agatino Battaglia, Tom Berney, Nadia Bolshakova, Sven Bölte, Patrick F. Bolton, Thomas Bourgeron, Sean Brennan, Jessica Brian, Susan E. Bryson, Andrew R. Carson, Guillermo Casallo, Jillian Casey, Brian H Y. Chung, Lynne Cochrane, Christina Corsello, Emily L. Crawford, Andrew Crossett, Cheryl Cytrynbaum, Geraldine Dawson, Maretha de Jonge, Richard Delorme, Irene Drmic, Eftichia Duketis, Frederico Duque, Annette Estes, Penny Farrar, Bridget A. Fernandez, Susan E. Folstein, Eric Fombonne, Christine M. Freitag, John Gilbert, Christopher Gillberg, Joseph T. Glessner, Jeremy Goldberg, Andrew Green, Jonathan Green, Stephen J. Guter, Hakon Hakonarson, Elizabeth A. Heron, Matthew Hill, Richard Holt, Jennifer L. Howe, Gillian Hughes, Vanessa Hus, Roberta Igliozzi, Cecilia Kim, Sabine M. Klauck, Alexander Kolevzon, Olena Korvatska, Vlad Kustanovich, Clara M. Lajonchere, Janine A. Lamb, Magdalena Laskawiec, Marion Leboyer, Ann Le Couteur, Bennett L. Leventhal, Anath C. Lionel, Xiao-Qing Liu, Catherine Lord, Linda Lotspeich, Sabata C. Lund, Elena Maestrini, William Mahoney, Carine Mantoulan, Christian R. Marshall, Helen McConachie, Christopher J. McDougle, Jane McGrath, William M. McMahon, Alison Merikangas, Ohsuke Migita,

Nancy J. Minshew, Ghazala K. Mirza, Jeff Munson, Stanley F. Nelson, Carolyn Noakes, Abdul Noor, Gudrun Nygren, Guiomar Oliveira, Katerina Papanikolaou, Jeremy R. Parr, Barbara Parrini, Tara Paton, Andrew Pickles, Marion Pilorge, Joseph Piven, Chris P. Ponting, David J. Posey, Annemarie Poustka, Fritz Poustka, Aparna Prasad, Jiannis Ragoussis, Katy Renshaw, Jessica Rickaby, Wendy Roberts, Kathryn Roeder, Bernadette Roge, Michael L. Rutter, Laura J. Bierut, John P. Rice, Jeff Salt, Katherine Sansom, Daisuke Sato, Ricardo Segurado, Ana F. Sequeira, Lili Senman, Naisha Shah, Val C. Sheffield, Latha Soorya, Inês Sousa, Olaf Stein, Nuala Sykes, Vera Stoppioni, Christina Strawbridge, Raffaella Tancredi, Katherine Tansey, Bhooma Thiruvahindrapduram, Ann P. Thompson, Susanne Thomson, Ana Tryfon, John Tsiantis, Herman Van Engeland, John B. Vincent, Fred Volkmar, Simon Wallace, Kai Wang, Zhouzhi Wang, Thomas H. Wassink, Caleb Webber, Rosanna Weksberg, Kirsty Wing, Kerstin Wittemeyer, Shawn Wood, Jing Wu, Brian L. Yaspan, Danielle Zurawiecki, Lonnie Zwaigenbaum, Joseph D. Buxbaum, Rita M. Cantor, Edwin H. Cook, Hilary Coon, Michael L. Cuccaro, Bernie Devlin, Sean Ennis, Louise Gallagher, Daniel H. Geschwind, Michael Gill, Jonathan L. Haines, Joachim Hallmayer, Judith Miller, Anthony P. Monaco, John I Nurnberger, Jr, Andrew D. Paterson, Margaret A. Pericak-Vance, Gerard D. Schellenberg, Peter Szatmari, Astrid M. Vicente, Veronica J. Vieland, Ellen M. Wijsman, Stephen W. Scherer, James S. Sutcliffe, and Catalina Betancur. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304):368–372, Jul 2010.

[16] Chiara Sabatti, Susan K Service, Anna-Liisa Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G Jones, Noah A Zaitlen, Teppo Varilo, Marika Kaakinen, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35–46, 2008.

[17] Pak C. Sham and Shaun M. Purcell. Statistical power and significance testing in large-

scale genetic studies. *Nat Rev Genet*, 15(5):335–346, May 2014.

[18] Kai Wang, Mingyao Li, and Hakon Hakonarson. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*, 11(12):843–854, Dec 2010.

[19] Lingjie Weng, Fabio Macciardi, Aravind Subramanian, Guia Guffanti, Steven G. Potkin, Zhaoxia Yu, and Xiaohui Xie. Snp-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics*, 12:99, 2011.

[20] Zhiyuan Xu, Xiaotong Shen, Wei Pan, Alzheimer's Disease Neuroimaging Initiative, et al. Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. *PloS one*, 9(8):e102312, 2014.

[21] Kenny Q. Ye and Corinne D. Engelman. Detecting multiple causal rare variants in exome sequence data. *Genet Epidemiol*, 35 Suppl 1:S18–S21, 2011.