Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Backward genotype-transcript-phenotype association mapping

Seunghak Lee [a,*], Haohan Wang [b], Eric P. Xing [b]

[a] Human Longevity, Inc., United States
[b] School of Computer Science, Carnegie Mellon University, United States

ARTICLE INFO

Article history:
Received 16 February 2017
Received in revised form 24 August 2017
Accepted 5 September 2017
Available online 14 September 2017

ABSTRACT

Genome-wide association studies have discovered a large number of genetic variants associated with complex diseases such as Alzheimer's disease. However, the genetic background of such diseases is largely unknown due to the complex mechanisms underlying genetic effects on traits, as well as a small sample size (e.g., 1000) and a large number of genetic variants (e.g., 1 million). Fortunately, datasets that contain genotypes, transcripts, and phenotypes are becoming more readily available, creating new opportunities for detecting disease-associated genetic variants. In this paper, we present a novel approach called "Backward Three-way Association Mapping" (BTAM) for detecting three-way associations among genotypes, transcripts, and phenotypes. Assuming that genotypes affect transcript levels, which in turn affect phenotypes, we first find transcripts associated with the phenotypes, and then find genotypes associated with the chosen transcripts. The backward ordering of association mappings allows us to avoid a large number of association testings between all genotypes and all transcripts, making it possible to identify three-way associations with a small computational cost. In our simulation study, we demonstrate that BTAM significantly improves the statistical power over "forward" three-way association mapping that finds genotypes associated with both transcripts and phenotypes and genotype-phenotype association mapping. Furthermore, we apply BTAM on an Alzheimer's disease dataset and report top 10 genotype-transcript-phenotype associations.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

One of fundamental problems in genetics is to find associations between genetic variants and phenotypes (e.g., disease status). In the past decade, researchers have detected associated single nucleotide polymorphisms (SNPs) with various diseases and traits such as Alzheimer's disease (AD) and height. However, it still remains a challenge to find genetic variants that influence complex diseases caused by multiple genetic and environmental factors [8]. Because of advancements in genome sequencing, it is becoming more affordable to generate rich datasets containing genotypes, transcripts, and phenotypes. To detect three-way associations among SNPs, transcripts, and phenotypes, various methods have been developed including NETAM [7] and a visual analytics approach with GenAMap [2]. NETAM is a network-driven method that constructs a network, where genotypes, transcripts, and phenotypes are nodes and associations between a pair of nodes are edges. It then finds paths from SNPs to phenotypes in the network, where the paths represent genotype-transcript-phenotype associations. GenAMap provides a visual analytics tool for biologists to investigate three-way associations given associations between genotypes and transcripts and between transcripts and phenotypes. Huang et al. [6] presented a novel strategy for three-way association mapping, where we find SNPs associated with phenotypes and among the chosen SNPs, we select SNPs associated with gene expressions. After that, the relationships between the chosen genes and the phenotypes are examined. We call it "forward" three-way association mapping to contrast it with our proposed approach of Backward Three-way Association Mapping (BTAM).

Furthermore, gene-based association mapping methods such as integrative network analysis [10], PrediXcan [3], MetaXcan [1], and transcriptome-wide association study (TWAS) [4] have been developed to utilize all SNPs, gene expressions, and phenotypes. Integrative network analysis attempts to find subnetworks of causal genes whose expressions differ between case and control groups, where causal genes are found based on associations between SNPs and gene expressions. To identify genes associated with phenotypes using the datasets with only genotypes and phenotypes, PrediXcan and TWAS predict gene expressions from the genotypes using a separate dataset with genotypes and transcripts. Then they find associations between the predicted gene expressions and the phenotypes. MetaXcan and TWAS predict phenotype-associated genes directly from summary statistics for genotype-phenotype

* Corresponding author.
E-mail address: leeseunghak@gmail.com (S. Lee).

https://doi.org/10.1016/j.ymeth.2017.09.004
1046-2023/© 2017 Elsevier Inc. All rights reserved.

associations. Unlike these gene-based association analysis, the focus of BTAM is to detect SNPs associated with phenotypes via transcripts. It would be possible to apply BTAM to the datasets with only genotypes and phenotypes if one can predict gene expressions from genotypes; however, this is beyond the scope of this paper and left for future work.

We start with an assumption that SNPs affect the phenotypes via expressed transcripts (e.g., gene expressions). Under the assumption, in this paper, we present a novel approach, referred to as BTAM, that first finds transcripts associated with phenotypes and then finds SNPs associated with the selected transcripts in the previous step. We call the direction for our three-way association analysis backward because it is opposite to the direction of information flow from SNPs to phenotypes. Fig. 1 shows the schematic diagram of (a) BTAM and (b) forward three-way association mapping [6] with arrows indicating the genetic information flow. BTAM consists of two steps. In the first step, we detect transcripts associated with phenotypes, and in the second step, we find SNPs associated with the survived transcripts of the first step. The backward direction of BTAM is motivated to test associations between all SNPs and the phenotype-associated transcripts, rather than all transcripts. In a comparison between backward and forward three-way association mapping, the former detects phenotype-associated SNPs via transcripts while the latter finds SNPs associated with both transcripts and phenotypes. In our simulation study, we demonstrate that BTAM achieves better statistical power over two-way genotype-phenotype association mapping and forward three-way association mapping. Finally, we apply BTAM to the AD dataset, and report top 10 SNP-transcript-AD status associations with some biological analyses via a literature review.

2. Methods

2.1. BTAM: Backward genotype-transcript-phenotype association mapping

Our goal is to find associations between genotypes $\mathbf{X} \in \mathbb{R}^{N \times J}$ and phenotypes $\mathbf{Z} \in \mathbb{R}^{N \times M}$, taking advantage of transcripts $\mathbf{Y} \in \mathbb{R}^{N \times K}$, where N is the sample size, J is the number of SNPs, K is the number of transcripts, and M is the number of phenotypes. Given a matrix $\mathbf{X}$, we denote the j-th column by $\mathbf{x}_j$, the i-th row by $\mathbf{x}^i$, and the (i,j) element by x_j^i. We define the model to generate phenotypes from genotypes as follows:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}_1, \quad (1)$$

$$\mathbf{Z} = \sigma(\mathbf{YG} + \mathbf{E}_2), \quad (2)$$

where $\mathbf{B}$ and $\mathbf{G}$ are the matrices, whose nonzero coefficients encode associations between $\mathbf{X}$ and $\mathbf{Y}$ and between $\mathbf{Y}$ and $\mathbf{Z}$, $\sigma(\mathbf{T})$ is a sigmoid function applied on each element in $\mathbf{T}$, and $\mathbf{E}_1$ and $\mathbf{E}_2$ are independent and identically distributed noise terms that follow Gaussian with zero mean and the identity covariance matrix. Recall that we consider the scenario where SNPs affect phenotypes via transcripts (i.e., gene expressions). Thus detecting SNPs that directly affect phenotypes is not the focus of this paper.

We describe BTAM assuming that $M = 1$ because each phenotype can be analyzed independently. In the first step (step1 in Fig. 1(a)), BTAM attempts to find transcripts associated with a phenotype. To achieve this, we adopt "screen and clean" strategy [13]. Here, "screen" refers to the step to discard irrelevant input variables using a multivariate regression method and "clean" refers to the step for statistical testing on the survived input variables. In the screening step, we first run ridge regression [5] on the transcripts as inputs and the phenotype as an output and select the transcripts whose absolute values of coefficients are greater than the mean of the absolute values of all coefficients. For the testing step, we run statistical tests (e.g., Wald test with logistic regression) between the survived transcripts and the phenotype. Given the p-values for transcripts, we choose a set of transcripts $\{S_m\}$ at the level of α_1 (e.g., $\alpha_1 = 0.05$) under false discovery rate (FDR) control (e.g., Benjamini-Hochberg procedure) or Bonferroni correction.

In the second step, given $\{S_m\}$ and genotypes, we run the screen and clean again on the genotype and the transcript data (step2 in Fig. 1(a)). Unlike the first step, we use lasso instead of ridge regression to select SNPs because lasso is effective for feature selection on very high dimensional data [15]. We note that due to the characteristics of ℓ_1 penalty employed by lasso, we may randomly choose one among highly correlated multiple SNPs with associations. Thus, to search for true causal SNPs, we need to investigate SNPs highly correlated with the SNPs found by BTAM (e.g., closely located SNPs in the genome) in a downstream analysis. Furthermore, we use a univariate linear regression test (e.g., F-test) for mapping between the survived transcripts and the phenotype because transcript data are continuous. For each SNP $\mathbf{x}_j$, the statistical test produces K' p-values, where K' is the number of survived transcripts. We assign the minimum of K' p-values to each SNP to avoid missing any significant associations. Finally, we determine the set of association SNPs under FDR control or Bonferroni correction at the level of α_2 (e.g., $\alpha_2 = 0.05$).

The backward ordering of BTAM has a better computational complexity than the reverse ordering of three-way association mapping (i.e., genotype-transcript mapping followed by transcript-phenotype mapping). The time complexity of BTAM is $O(KM + JK')$, where K' is the number of transcripts associated with

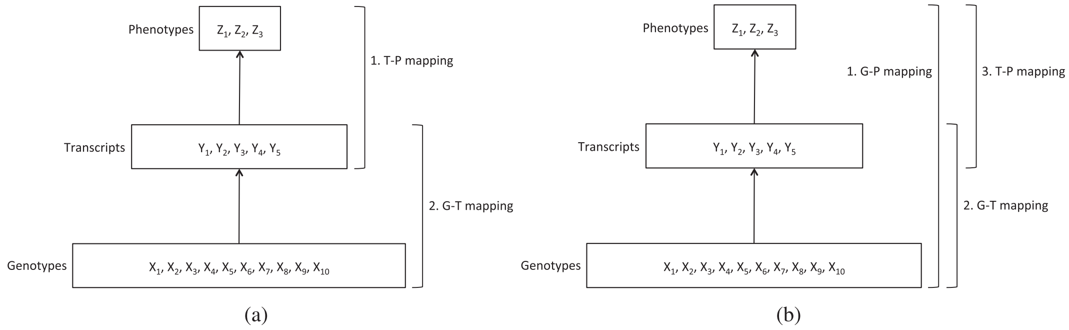
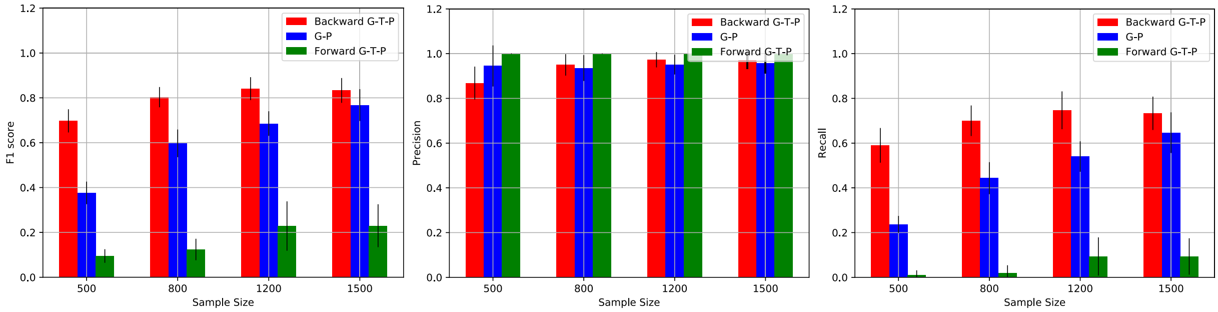
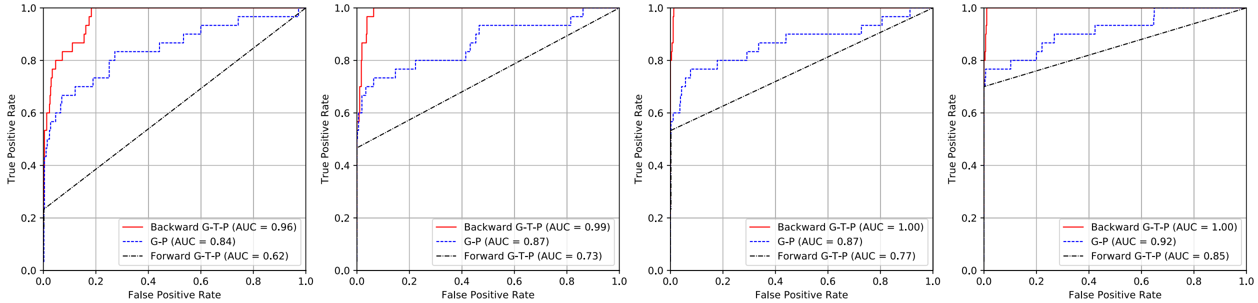


Fig. 1. Schematic diagram of (a) backward three-way association mapping (BTAM) and (b) forward three-way association mapping [6]. In (a), we first find transcripts associated with phenotypes, and then find SNPs associated with the transcripts with any associations with the phenotypes. In (b), we first find SNPs associated with phenotypes, and then among the SNPs associated with the phenotypes, we select SNPs associated with at least one transcript. We then investigate relationships between the transcripts and the phenotypes.



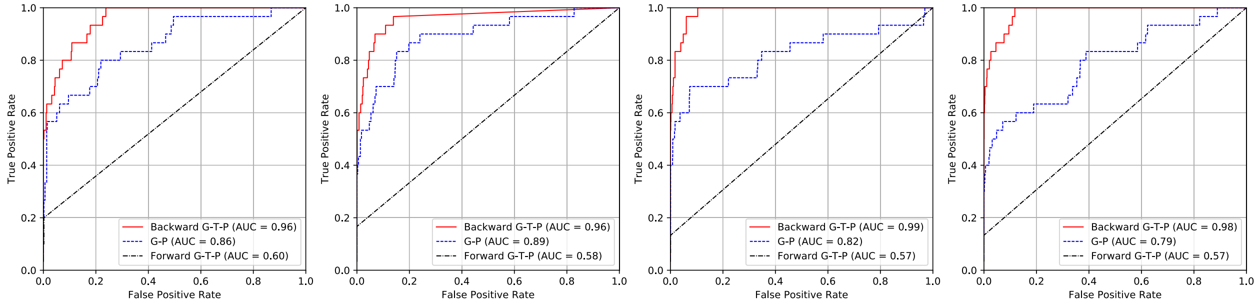
(a) (b) (c)

Fig. 2. Simulation results of backward three-way (Backward G-T-P), forward three-way (Forward G-T-P), and two-way (G-P) association mapping in terms of (a) F1 score, (b) precision, and (c) recall rates under different sample sizes from 500 to 1500. Here G, T, and P represent genotype, transcript, and phenotype, respectively. Average performances on 10 repetitions are shown with error bar of one standard deviation.



(a) (b) (c) (d)

Fig. 3. ROC curves for the simulation results of backward three-way (Backward G-T-P), forward three-way (Forward G-T-P), and two-way (G-P) association mapping with different sample sizes of (a) $N = 500$, (b) $N = 800$, (c) $N = 1200$, and (d) $N = 1500$. Here G, T, and P represent genotype, transcript, and phenotype, respectively. AUC (Area Under the Curve) is shown in the legend.



(a) (b) (c) (d)

Fig. 4. ROC curves for the simulation results of backward three-way (Backward G-T-P), forward three-way (Forward G-T-P), and two-way (G-P) association mapping with different number of SNPs of (a) $J = 500$, (b) $J = 1000$, (c) $J = 1500$, and (d) $J = 2000$. Here G, T, and P represent genotype, transcript, and phenotype, respectively. AUC (Area Under the Curve) is shown in the legend.

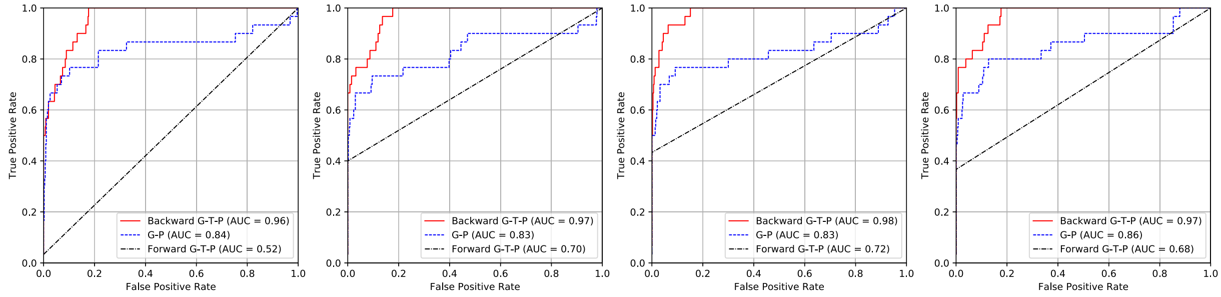
at least one phenotype. In comparison, the time complexity of three-way association mapping with the reverse ordering is $O(K''M + JK)$, where K'' is the number of transcripts having at least one association SNP. Typically, $J \gg M, K$ and thus the former is significantly faster than the latter.

We note that in association mappings, it is important to control confounding effects such as ancestry and age that lead false discoveries. A simple approach to correct for known confounding effects is to add them as covariates into a linear or logistic regression model. In BTAM, we include covariates into the models for all screen and clean steps. In our AD data analysis, as covariates, we included the age of death and top 10 principal components from transcripts in step 1, and the age of death and top 10 principal components from genotypes in step 2.

3. Results

3.1. Simulation study

We demonstrate that BTAM improves the statistical power over forward three-way association mapping (FTAM) [6] and two-way genotype-phenotype association mapping using Wald test with logistic regression. For FTAM, we run step 1 and 2 in Fig. 1(b) and skip step 3 because the purpose of our simulations is to identify phenotype-associated SNPs. We generated simulation datasets of genotypes, transcripts, and one balanced phenotype of disease status (case:1, control:0). We set the number of phenotypes to one because multiple phenotypes can be processed independently. We first created SNPs under Hardy–Weinberg equilibrium with a



(a) (b) (c) (d)

Fig. 5. ROC curves for the simulation results of backward three-way (Backward G-T-P), forward three-way (Forward G-T-P), and two-way (G-P) association mapping with different minor allele frequencies of (a) 0.05, (b) 0.1, (c) 0.15, and (d) 0.2 for association SNPs. Here G, T, and P represent genotype, transcript, and phenotype, respectively. AUC (Area Under the Curve) is shown in the legend.

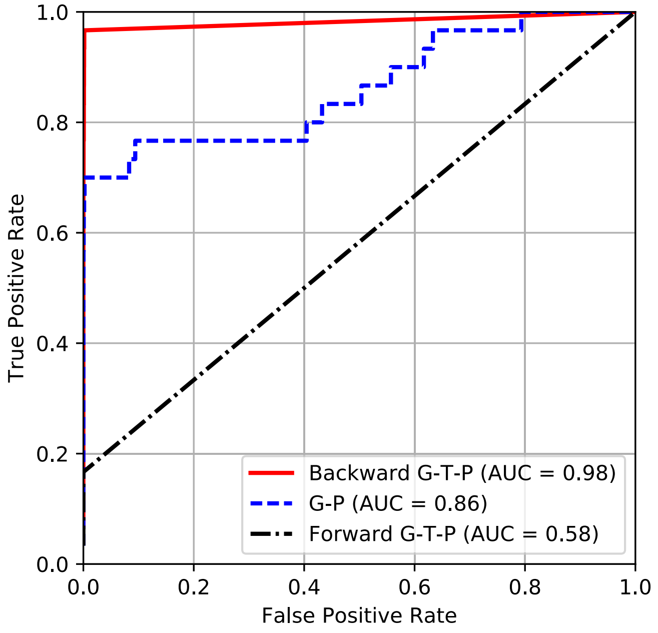


Fig. 6. ROC curve for the simulation results of backward three-way (Backward G-T-P), forward three-way (Forward G-T-P), and two-way (G-P) association mapping under the setting of 1000 samples, 500000 SNPs, 20000 transcripts, and a phenotype. Here G, T, and P represent genotype, transcript, and phenotype, respectively. AUC (Area Under the Curve) is shown in the legend.

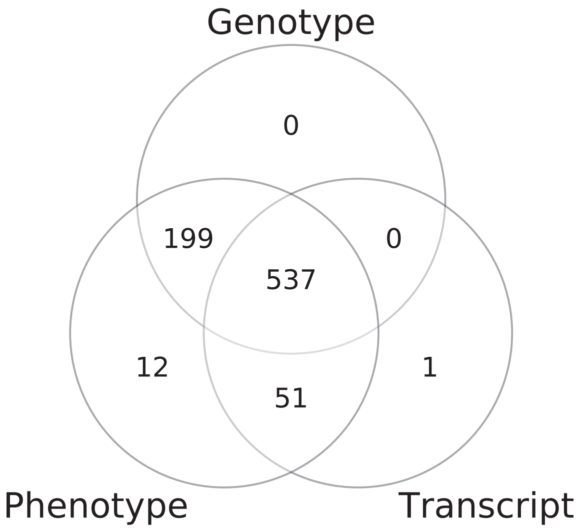


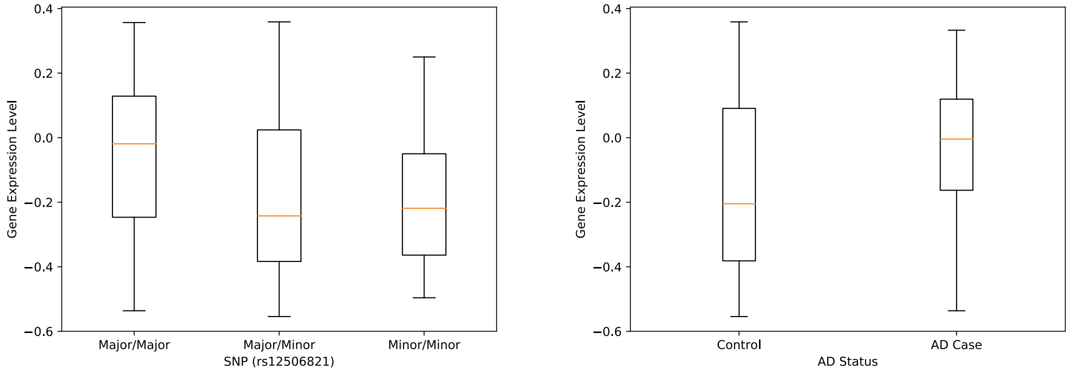
Fig. 7. Venn diagram representing the number of individuals containing genotypes, transcripts, and the AD status phenotype in the AD dataset [14].

fixed minor allele frequency (MAF). Then, the transcript levels were generated by $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, where $\mathbf{B}$ encodes the effects of SNPs on the transcript levels and $\mathbf{E}$ is Gaussian noise with zero mean and the identity covariance matrix. For the phenotypes, we computed $\mathbf{Yg} + \mathbf{E}$ and selected the individuals with $\lfloor N/2 \rfloor$ largest values as cases and the other half as controls, where $\mathbf{g}$ contains the fixed effects of transcripts on the phenotype. Furthermore, as a covariate, we used a random variable defined by $0.5\mathbf{Z} + 0.5\mathbf{E}$. In this simulation study, we tested the methods in both small-scale and large-scale settings to demonstrate the behavior of BTAM under various settings and the usefulness of BTAM under large-scale genome-wide association studies. In the small-scale setting, we set the number of samples to 500, the number of SNPs to 500, the number of genes to 50, the number of association SNPs to 30, and the number of association genes to 3, and draw the effect sizes of SNPs from $\mathcal{N}(1,1)$ and the effect sizes of genes from $\mathcal{N}(2,1)$. Then we report the simulation results by changing the number of samples, the number of SNPs, and MAFs. The large-scale setting is the same as the small one except that we fixed the number of individuals, SNPs, and transcripts to 1000, 500000, and 20000, respectively. Then, for the large-scale experiments, we report both performance and runtime. For BTAM, we set $\alpha_1 = \alpha_2 = 0.05$ to obtain the results in Fig. 2. When generating Receiver Operating Characteristic (ROC) curves in Figs. 3–5 we set $\alpha_1 = 0.05$ and did not use FDR control after the second step. To determine the regularization parameter for ridge regression, we used leave-one-out cross validation over $\{0.1, 1.0, 10.0\}$. For lasso, it was chosen using 2-fold cross validation over $\{0.01, 0.05, 0.1\}$. In the cross validations, parameters were evaluated by the corresponding prediction errors. To make a fair comparison, we applied the same level of FDR control on all methods.

We first report the results in small-scale settings. Fig. 2 shows the average performances of BTAM (Backward G-T-P), FTAM (Forward G-T-P), and two-way association mapping (G-P) over 10 repeated simulations with different sample sizes from 500 to 1500 measured in F1 score, precision, and recall rates. Here, G, T, and P represent genotype, transcript, and phenotype. BTAM outperformed the other two methods in terms of F1 score and recall; however, FTAM was the best in terms of precision. It verifies that BTAM can take advantage of transcripts to finds more association SNPs than FTAM and two-way association mapping. While FTAM showed low recall rate, its precision was high because FTAM found SNPs associated with both phenotypes and transcripts. In Figs. 3–5, we show the simulation results in terms of ROC curves for BTAM,

Table 1
Top 10 SNP-transcript-AD status associations, identified by BTAM. G-T pval and T-P pval represent p-values for associations between the corresponding SNP and the transcript and between the corresponding transcript and the AD status, respectively.

SNP	SNP Location	Transcript/Gene	Gene Location	G-T pval	T-P pval
rs10027926	chr4:3412927	*HSS00209220*		9.94×10^{-12}	6.82×10^{-8}
rs3129317	chr4:3285928	*HSS00209220*		2.93×10^{-10}	6.82×10^{-8}
rs16844383	chr4:3445516	*HSS00209220*		1.29×10^{-9}	6.82×10^{-8}
rs12506821	chr4:3282833	*DLG4*	chr17:7189890-7220050	2.11×10^{-8}	1.08×10^{-7}
rs3113773	chr4:180619604	*HSS00209220*		7.53×10^{-8}	6.82×10^{-8}
rs4669778	chr2:11743854	*BDNF*	chr11:27654893-27722058	2.56×10^{-7}	2.05×10^{-8}
rs10513502	chr3:156984072	*HOOK2*	chr19:12763003-12872740	4.59×10^{-7}	4.45×10^{-9}
rs1144962	chr12:68941205	*BDNF*	chr11:27654893-27722058	4.88×10^{-7}	1.19×10^{-7}
rs1955452	chr14:28734495	*hCT2326510*		5.5×10^{-7}	1.32×10^{-7}
rs10784735	chr12:68953091	*BDNF*	chr11:27654893-27722058	5.87×10^{-7}	1.19×10^{-7}



(a)

(b)

Fig. 8. Relationships (a) between rs12506821 and the expression level of transcript *DLG4* and (b) between AD status and the expression level of transcript *DLG4*.

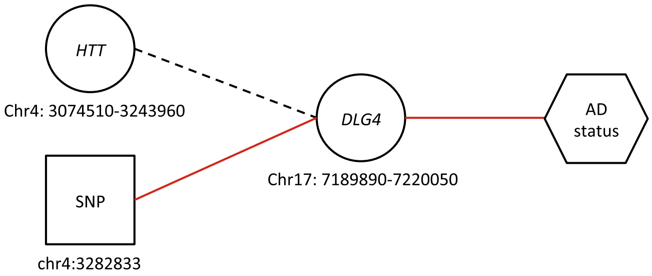


Fig. 9. An example of three-way association found by BTAM that includes rs12506821 (chr4:3282833), *DLG4* (chr17:7189890-7220050), and AD status. The gene *HTT* (chr4: 3074510-3243960) interacts with *DLG4* [12]. The solid red lines represent associations found by BTAM and the dotted line represents a gene-gene interaction.

FTAM, and two-way association mapping. It shows that BTAM is better than the other alternatives under various settings, demonstrating that BTAM is a promising method for three-way association analysis.

Fig. 6 shows the results under a large-scale setting with 1000 samples, 500000 SNPs, 20000 transcripts, and a phenotype. In the case, BTAM also significantly outperformed the other methods. We note that the performance gap between three-way association mappings (BTAM and FTAM) and two-way association mapping was very large. Thus, it suggests that three-way association mapping is a promising approach to genome-wide association studies.

To test the computational efficiency, we measured the runtime for the large-scale simulations over three different runs. On average, BTAM took 7 min, FTAM took 60 min, and two-way association mapping took 57 min. More specifically, BTAM used 1 min and 6 min for transcript-phenotype and genotype-transcript mapping, respectively. It confirms that BTAM is not only more statistically powerful but also more computationally efficient than the two alternatives.

3.2. Association analysis of Alzheimer's disease dataset

We applied BTAM on late-onset AD dataset [14] generated by Harvard and Merck Pharmaceutical and supplied through the Sage Bionetworks Repository. We used an AD cohort matched for age, gender, and post mortem interval that involves genotypes, transcript levels from the cerebellum (profiled on custom-made Agilent 44K microarray, which measures expression levels of 37585 known and predicted genes, miRNAs and non-coding RNAs), and the phenotype for AD status. As a quality control on the data, we excluded four individuals whose first or second principal component from the genomes is > 10 standard deviations away from the mean. The cohort in this analysis contains 736 individuals with 555091 SNPs, 589 individuals with 40638 transcripts, and 799 individuals with AD status. The number of individuals with both genotypes and transcripts is 537 and the number of individuals with both transcripts and phenotypes is 588. Fig. 7 shows a Venn diagram for the number of individuals with the genotypes, the transcripts, and the phenotype of AD status. We imputed the missing values using the mean of the corresponding SNP or transcript.

We applied BTAM to the AD dataset consisting of genotypes, transcripts, and AD status. In summary, we found 15 transcripts for transcript-phenotype association mapping with $\alpha_1 = 0.01$ after Bonferroni correction, and found 2703 association SNPs with $\alpha_2 = 0.01$ under FDR control. To determine the regularization parameters for ridge regression and lasso, we used the cross validation settings in Section 3.1. Table 1 shows top 10 three-way associations identified by BTAM, sorted by p-values for genotype-transcript associations.

As an example of association analysis, we investigate the biological underpinnings of a three-way association that involves rs12506821, *DLG4*, and AD status, where p-values for rs12506821-*DLG4* and *DLG4*-AD status associations are 2.11×10^{-8} and 1.08×10^{-7}. Both p-values are statistically significant at the level of 0.05 after Bonferroni correction. The gene *DLG4* (Discs Large MAGUK Scaffold Protein 4) is located in a protein coding region (www.genecards.org) and is associated with Alzheimer's disease [9]. It has been reported that interaction between *DLG4* and *Fyn* affects chronic NMDAR hyperactivity in AD, which leads to the reduced level of the proteins that support neuronal survival such as *BDNF* (Brain Derived Neurotrophic Factor). As a result, it may cause neuronal death [9]. Note that in Table 1, three associations (out of 10) involve *BDNF* that plays an important role in neuronal survival or growth. Fig. 8(b) shows that *DLG4* is highly expressed in AD patients compared to the control individuals, and Fig. 8(a) shows that a minor allele in rs12506821 reduces the expression level of *DLG4*. In our literature review, we found out that *HTT* (Huntington) is a gene interacting with *DLG4* [12], and *HTT* is located 38873bp upstream of rs12506821 [11]. Fig. 9 is an illustration of the three-way association. Based on the observations, a hypothesis for the three-way association would be that rs12506821 has a cis-effect on *HTT* and *HTT* influences on *DLG4* through interactions. Consequently, the perturbed expression levels of *DLG4* affect the risk of AD. We note that the hypothesis needs to be confirmed by independent biological studies.

4. Conclusions

We presented a novel method called BTAM for detecting genotype-transcript-phenotype association mapping in a backward direction such that transcript-phenotype association mapping is performed prior to genotype-transcript association mapping. Backward association mapping is computationally beneficial because it allows us test associations between all SNPs against a set of transcripts associated with a phenotype of interest, rather than all transcripts. In our simulations, we showed that BTAM achieves significantly better statistical power compared to other alternatives. Furthermore, we analyzed the AD dataset using BTAM and reported top 10 SNP-transcript-AD status associations. We also investigated an three-way association via a literature review and discovered some evidence for their biological relevance to AD.

A promising future direction of research would be to integrate multiple datasets in different layers under the framework of BTAM. For example, we may add clinical phenotypes and disease status layer on top of transcript layer. Furthermore, to improve the statistical power of BTAM, in transcript layer we may include multiple datasets such as transcripts from different parts of brain (e.g., prefrontal cortex, visual cortex) and perform multi-trait association mapping between genotypes and transcripts. Another interesting direction would be to investigate epistatic effects of multiple SNPs on AD status via a shared association transcript. If there exist nonlinear associations between the expression level of the shared transcript and AD status, it would be possible to find non-additive effects of multiple SNPs on AD status.

Acknowledgements

This work was done under a support from NIH R01GM114311 and NIH P30DA035778.

References

[1] Alvaro Barbeira et al., MetaXcan: summary statistics based gene-level association method infers accurate PrediXcan results, BioRxiv, 2016, http://dx.doi.org/10.1101/045260.

[2] Ross E. Curtis, Junming Yin, Peter Kinnaird, Eric P. Xing, Finding genome-transcriptome-phenome association with structured association mapping and visualization in GenAMap, in: Pacific Symposium on Biocomputing, 2012, pp. 327–338.

[3] Eric R. Gamazon et al., A gene-based association method for mapping traits using reference transcriptome data, Nat. Genet. 47 (9) (2015) 1091–1098.

[4] Alexander Gusev et al., Integrative approaches for large-scale transcriptome-wide association studies, Nat. Genet. 48 (3) (2016) 245–252.

[5] Arthur E. Hoerl, Robert W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics (1970) 55–67.

[6] R. Stephanie Huang et al., A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity, Proc. Nat. Acad. Sci. 104 (23) (2007) 9758–9763.

[7] Seunghak Lee, Soonho Kong, Eric P. Xing, A network-driven approach for genome-wide association mapping, Bioinformatics 32 (12) (2016) i164–i173.

[8] Teri A. Manolio et al., Finding the missing heritability of complex diseases, Nature 461 (7265) (2009) 747–753.

[9] E. Mohandas, V. Rajmohan, B. Raghunath, Neurobiology of Alzheimer's disease, Indian J. Psychiatry 51 (1) (2009) 55.

[10] Eric E. Schadt et al., An integrative genomics approach to infer causal associations between gene expression and disease, Nat. Genet. 37 (7) (2005) 710–717.

[11] Stephen T. Sherry et al., dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (1) (2001) 308–311.

[12] Damian Szklarczyk et al., STRING v10: protein–protein interaction networks, integrated over the tree of life, Nucleic Acids Res. 43 (D1) (2014) D447–D452.

[13] Jing Wu, Bernie Devlin, Steven Ringquist, Massimo Trucco, Kathryn Roeder, Screen and clean: a tool for identifying interactions in genome-wide association studies, Genet. Epidemiol. 34 (3) (2010) 275–285.

[14] Bin Zhang et al., Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease, Cell 153 (3) (2013) 707–720.

[15] Peng Zhao, Bin Yu, On model selection consistency of lasso, J. Mach. Learn. Res. 7 (2006) 2541–2563.