Cascading Epigenomic Model for GWAS

Bernard Ng[1,2], **William Casazza**[1,2], **Farnush Farhadi**,[1,2] **Sara Mostafavi**[1,2]
[1] Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z4
[2] Centre for Molecular Medicine and Therapeutics, Vancouver, British Columbia, Canada, V5Z 4H4
bernardyng@gmail.com

Abstract

Deciphering the mechanisms through which GWAS SNPs affect phenotypes is challenging. A number of methods, such as MetaXcan, have been put forth for associating the genetic component of either gene expression or DNA methylation to phenotypes. In this work, we propose a cascading epigenomic analysis for GWAS (CEWAS) to associate the epigenomic component of gene expression that is driven by genetics to phenotypes. On a well-powered GWAS, we show that CEWAS provides higher detection sensitivity than MetaXcan. Importantly, we show with simulations that statistics generated by CEWAS are properly calibrated. Further, using atSNP, we show that many SNPs associated with the detected genes affect transcription factor (TF) affinity. In fact, some of the detected genes are found to be potential TF targets, illustrating another utility of CEWAS.

1 Introduction

Genome wide association studies (GWAS) have identified thousands of SNPs related to complex traits and disease risk. Majority of these SNPs lies in non-protein coding regions, hence determining the mechanisms through which these SNPs act on phenotypes is nontrivial. Many studies use gene expression data to functionally map SNPs to genes, with expression quantitative trait locus (eQTL) analysis [1] being the most common approach. However, this approach suffers from the limitation that linkage disequilibrium (LD) can result in coincidental overlaps between eQTL and GWAS SNPs. To handle LD, a number of colocalization methods [2-4], such as summary data based Mendelian Randomization (SMR), have been proposed to identify pleiotropic SNPs that affect both gene expression and phenotypes. Alternatively, one can extract the genetic component of gene expression and associate that with phenotypes. Methods under this category includes PrediXcan [5], as well as summary statistics-based extensions, such as TWAS [6], MetaXcan [7], and MultiXcan [8].

Since GWAS SNPs are enriched in enhancers and open chromatin regions [9], their effects on phenotypes are likely exerted via gene regulation rather than directly on protein sequence modification [10]. By far, DNA methylation (DNAm) is the most widely-used epigenomic mark for examining gene regulation. Motivated by how associations between DNAm and SNPs (i.e. mQTLs) are seen genome-wide [11], a number of recent studies combine mQTL with GWAS using SMR [12-14] to investigate the epigenomic effects of GWAS SNPs. To reconcile with eQTL-based results, some studies further apply SMR to map CpGs to genes by finding CpG-gene pairs with shared genetic effects [12; 14].

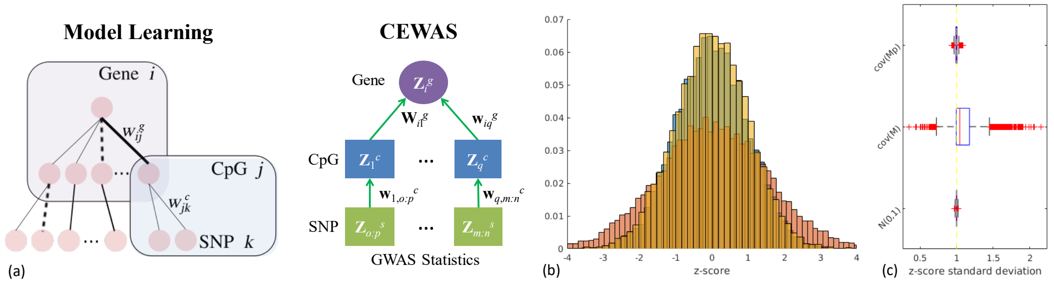

Fig. 1: CEWAS and z-score calibration. (a) CEWAS entails first building a set of models for predicting gene expression and DNAm levels. These models are then applied to GWAS summary statistics, $\mathbf{z}_k{}^s$, to estimate gene level z-scores, $\mathbf{z}_i{}^g$. (b) The statistics, $\mathbf{z}_i{}^g$, (yellow) for an exemplar gene generated by applying CEWAS to 10000 sets of null $\mathbf{z}_k{}^s$ follow a standard normal distribution (blue), confirming that $\mathbf{z}_i{}^g$ is properly calibrated. Using covariance estimated from the original DNAm levels, $\mathbf{M}$, as opposed to only the genetic component of DNAm, $\mathbf{M}_\text{p}$, inflates $\mathbf{z}_i{}^g$ (red). (c) Standard deviation (std) of 10000 $\mathbf{z}_i{}^g$'s for each gene i shown across all genes. The std of $\mathbf{z}_i{}^g$ (with $\mathbf{M}_\text{p}$) is ~1 for all genes.

An immediate question is whether we can directly associate the epigenomic component of gene expression that is driven by genetics to phenotypes, as opposed to analyzing all data-type pairs and trying to tie results together heuristically. For this, we propose a cascading epigenomic analysis for GWAS, which we refer to as CEWAS (Fig. 1a). With MetaXcan as the building block, the idea is to build a prediction model for extracting the epigenomic component of expression for each gene and a corresponding set of prediction models for extracting the genetic component of DNAm levels for CpGs proximal to the given gene. Applying these models to GWAS summary statistics provides gene-level z-scores that reflect genetically-driven epigenomic effects. To test this approach, we first build the respective prediction models using imputed genotype, DNAm, and RNAseq data from the ROSMAP study [15]. We then apply CEWAS to a well-powered GWAS [16] and show that it provides higher detection power compared to MetaXcan. Importantly, we confirm with simulations that CEWAS produces calibrated z-scores. Moreover, we compare the SNP sets associated with genes detected by CEWAS and MetaXcan in terms of their impacts on transcription factor (TF) binding affinity using atSNP [17]. More SNPs from the CEWAS set are found to affect TF binding affinity. In fact, some of the detected genes are shown to be potential TF targets.

2 Methods

2.1 Cascading Epigenomic Analysis for GWAS

Motivated by the observation that GWAS SNPs are enriched in regulatory regions [9], we propose CEWAS (Fig. 1a) to analyze the cascading effects of genetics from epigenome to transcriptome and eventually the phenome. With MetaXcan as the building block, we first build a model for extracting the epigenomic component of expression for each gene i: $\mathbf{E}_i = \Sigma_{j\epsilon Ci}\mathbf{w}_{ij}{}^g\mathbf{M}_j + \boldsymbol{\varepsilon}_i{}^g$, where $\mathbf{E}_i$ is a n×1 vector containing the expression level of gene i from n subjects, $\mathbf{M}_j$ is a n×1 vector containing the DNAm levels of CpG j, and $\mathbf{w}_{ij}{}^g$ is the j^th element of a m_i×1 model weight vector, $\mathbf{w}_i{}^g$, to be estimated. C_i is the set of m_i CpGs within 1Mb from the transcription starting site (TSS) of gene i. Following MetaXcan, we estimate $\mathbf{w}_{ij}{}^g$ using elastic net regression by applying GLMNET [18] with its default settings.

To extract the genetic component of DNAm for each CpG j, we model $\mathbf{M}_j$ in a similar manner: $\mathbf{M}_j = \Sigma_{k\epsilon Sj}\mathbf{w}_{jk}{}^c\mathbf{S}_k + \boldsymbol{\varepsilon}_j{}^c$, where $\mathbf{S}_k$ is a n×1 vector containing the dosage of SNP k, $\mathbf{w}_{jk}{}^c$ is the k^th element of a l_j×1 model weight vector, $\mathbf{w}_j{}^c$, to be estimated with elastic net regression, and S_j is the set of l_j SNPs within 100Kb from the CpG j.

Given $\mathbf{w}_{ij}{}^g$ and $\mathbf{w}_{jk}{}^c$, genetically-driven epigenomic effects at gene level can be estimated in a manner analogous to sequentially applying MetaXcan:

$$\begin{aligned}\mathbf{z}_i^g = \Sigma_{j\in Ci}\mathbf{w}_{ij}^g\boldsymbol{\sigma}_j^c/\boldsymbol{\sigma}_i^g \cdot \mathbf{z}_j^c &= \Sigma_{j\in Ci}\mathbf{w}_{ij}^g\boldsymbol{\sigma}_j^c/\boldsymbol{\sigma}_i^g \cdot (\Sigma_{k\in Sj}\mathbf{w}_{jk}^c\boldsymbol{\sigma}_k^s/\boldsymbol{\sigma}_j^c \cdot \mathbf{z}_k^s) \\ &= \Sigma_{j\in Ci}\mathbf{w}_{ij}^g\Sigma_{k\in Sj}\mathbf{w}_{jk}^c\boldsymbol{\sigma}_k^s/\boldsymbol{\sigma}_i^g \cdot \mathbf{z}_k^s,\end{aligned} \tag{1}$$

where $\mathbf{z}_i^g$ is the z-score at gene level for gene *i,* $\mathbf{z}_j^c$ is the z-score at CpG level for CpG *j*, and $\mathbf{z}_k^s$ is the z-score at SNP level for SNP *k*. $\boldsymbol{\sigma}_i^g$ and $\boldsymbol{\sigma}_k^s$ are the variance of gene *i* and SNP *k*, respectively. A gene is declared significant if the corresponding p-value of its $\mathbf{z}_i^g$ is less than 0.05 with Bonferroni correction. A critical deviation of CEWAS from MetaXcan is the way in which $\boldsymbol{\sigma}_i^g$ needs to be estimated. Since only genetically-driven epigenomic effects are retained by the DNAm prediction models, we must estimate $\boldsymbol{\sigma}_i^g$ based only on the genetic component of DNAm. For this, we set $\boldsymbol{\sigma}_i^g$ to $\mathbf{w}_i^{g\mathrm{T}}cov(\mathbf{M}_\mathrm{p})\mathbf{w}_i^g$, where $\mathbf{M}_\mathrm{p}$ is a $n\times m_\mathrm{i}$ matrix containing predicted DNAm levels. As will be shown, estimating $\boldsymbol{\sigma}_i^g$ with $\mathbf{M}_\mathrm{p}$ is critical for $\mathbf{z}_i^g$ to be calibrated.

2.2 CEWAS z-score Calibration

For the detected genes to be reliable, we need to ensure that $\mathbf{z}_i^g$ generated by CEWAS is calibrated. In particular, applying (1) to null $\mathbf{z}_k^s$ should output null $\mathbf{z}_i^g$. To properly test whether this criterion is met, the choice of input to (1) is critical. In addition to requiring $\mathbf{z}_k^s$ of each SNP *k* to follow $N(0,1)$, the LD structure of all SNPs involved in (1), i.e. all nodes at the bottom layer in Fig. 1a, must be accounted for. To satisfy these two conditions, for each gene *i*, we draw 10000 sets of $\mathbf{z}_k^s$ from $N(\mathbf{0},\mathbf{R}_i)$, where $\mathbf{R}_i$ is the correlation between all SNPs in S_j for $j \in C_i$. Using correlation, as opposed to covariance, ensures the standard deviation of each $\mathbf{z}_k^s$ is 1. For generating $\mathbf{R}_i$, we use the ROSMAP imputed genotype data. For evaluation, we assess if each set of 10000 $\mathbf{z}_i^g$'s of each gene *i* follows $N(0,1)$.

2.3 TF and Gene Target Identification

We hypothesize that SNPs (those with $\mathbf{w}_{ij}^g\mathbf{w}_{jk}^c > 0$) of genes detected by CEWAS might be regulatory with a portion of their effects exerted via altering TF binding affinity. To test this hypothesis, we apply atSNP [17] to determine if any TFs (from the ENCODE project [19]) are affected by these SNPs. SNP-TF motif pairs are declared as significant at an α of 0.05 with Bonferroni correction for the number of pairs tested. For each detected SNP-TF motif pair, we further test if the associated gene (for which the SNP is given non-zero model weight) might be a TF target by applying a standard linear interaction model on the ROSMAP data. To test for any effect of TF *l* on gene expression, i.e. the main effect of TF *l* and its interaction with SNP *k* jointly, we apply extra sum of squares. Significance is declared at an α of 0.05 with Bonferroni correction for the number of SNP-TF-gene triples analyzed.

3 Materials

Imputed genotype, DNAm, and RNAseq data from the ROSMAP study [15] were used in this work. Genotype data were acquired from 2067 subjects. The DNAm and RNAseq data were derived from brain tissues of 702 and 698 subjects, respectively. 543 subjects have both genotype and DNAm data; 485 subjects have both DNAm and RNAseq data; and 534 subjects have both genotype and RNAseq data. Similar preprocessing pipelines were applied to the data as previously described [20].

4 Results and Discussion

We applied CEWAS models built from brain tissue data to a well-powered schizophrenia GWAS [16]. Although the ROSMAP subjects are of European descent, the reference allele for some SNPs could be different from the GWAS. We accounted for allele flips by inverting the sign of the GWAS z-scores, and removed all ambiguous SNPs, i.e. cases where A1 and A2 are complementary, e.g. A1=A, A2=T. For comparison, we applied MetaXcan with gene expression prediction models ($\mathbf{w}_{ik}{}^{g}$) built using the same sets of SNPs as CEWAS. We also built models with SNPs within 1Mb from TSS. Similar results were obtained hence not reported here.

Of the 10129 genes tested by CEWAS, 118 genes were found to be significant. In contrast, MetaXcan detected 89 genes out of 12525 genes tested. Hence, CEWAS seems to provide higher detection sensitivity, and this is true even when we used the same p-value threshold as MetaXcan. The difference in the number of tested genes was due to e.g. some genes having no proximal CpGs that were given non-zero model weights. The p-values between CEWAS and MetaXcan (restricting to those passing a nominal threshold of 0.05) are highly correlated (r=0.44, p<10^{-100}), suggesting some shared signals being detected. Importantly, our simulations (Section 2.2) confirmed that gene-level z-scores from CEWAS are calibrated (Fig. 1b&c), provided that we use predicted DNAm levels to estimate $\boldsymbol{\sigma}_i{}^{g}$ in (1). Hence, CEWAS's higher detection sensitivity compared to MetaXcan is not due to z-score inflation.

To decipher what gave rise to the differences in genes detected by CEWAS and MetaXcan, we assessed the SNPs (those with $\mathbf{w}_{ij}{}^{g}\mathbf{w}_{jk}{}^{c} > 0$ and $\mathbf{w}_{ik}{}^{g} > 0$, respectively) of the significant genes in terms of their effects on TF binding affinity. Using atSNP, we tested 32740575 SNP-TF motif pairs with 37262 pairs found to be significant for CEWAS (Fig. 2b), whereas 5577 out of 5086095 pairs are significant for MetaXcan. We suspect the increased number of significant SNP-TF motif pairs with CEWAS (p < 0.01 based on Fisher's exact test) might relate to interactions between DNAm and TF binding being indirectly captured by CEWAS. Also, 63 out of 118 genes detected by CEWAS are found to be potential TF targets. For example, GATAD2A is found to be associated with SP1 (Fig. 2c), and SP1 is a TF implicated in schizophrenia [21]. Thus, CEWAS can potentially be used for finding disease-related TF targets.

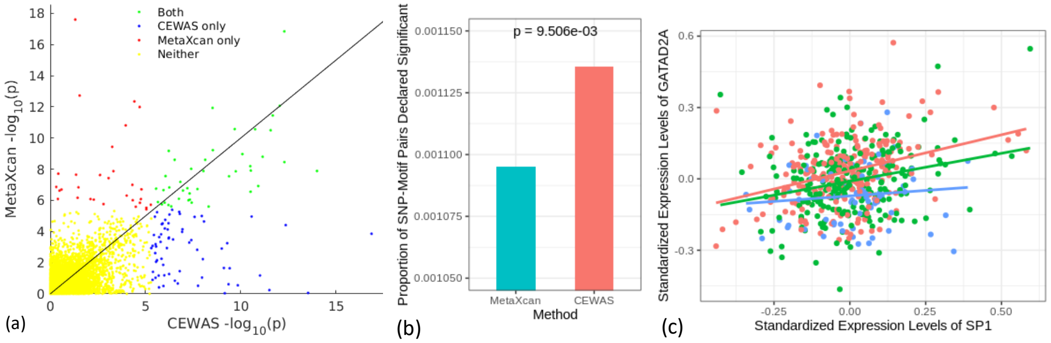

Fig. 2: CEWAS vs. MetaXcan. (a) Gene-level p-values of CEWAS vs. MetaXcan. (b) Proportion of SNP-TF motif pairs declared significant. (c) Standardized expression levels of SP1 vs. GATAD2A.

References

[1] Lappalainen, T., Sammeth, M., Friedlander, M. R., t Hoen, P. A., Monlong, J., Rivas, M. A., . . . Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature, 501*(7468), 506-511. doi: 10.1038/nature12531

[2] Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., & Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS genetics, 6*(4), e1000895.

[3] Wen, X., Pique-Regi, R., & Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS genetics, 13*(3), e1006646.

[4] Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., . . . Visscher, P. M. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics, 48*(5), 481.

[5] Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., . . . Cox, N. J. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics, 47*(9), 1091.

[6] Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., . . . Wright, F. A. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics, 48*(3), 245.

[7] Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., . . . Edwards, T. L. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature communications, 9*(1), 1825.

[8] Barbeira, A. N., Pividori, M. D., Zheng, J., Wheeler, H. E., Nicolae, D. L., & Im, H. K. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS genetics, 15*(1), e1007889.

[9] Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., . . . Coyne, M. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature, 473*(7345), 43.

[10] Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., . . . Brody, J. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science, 337*(6099), 1190-1195.

[11] Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T. M., . . . Mill, J. (2016). Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat Neurosci, 19*(1), 48-54. doi: 10.1038/nn.4182

[12] Hannon, E., Gorrie-Stone, T. J., Smart, M. C., Burrage, J., Hughes, A., Bao, Y., . . . Mill, J. (2018). Leveraging DNA-methylation quantitative-trait loci to characterize the relationship between Methylomic variation, gene expression, and complex traits. *The American Journal of Human Genetics, 103*(5), 654-665.

[13] Hannon, E., Weedon, M., Bray, N., O'Donovan, M., & Mill, J. (2017). Pleiotropic effects of trait-associated genetic variation on DNA methylation: utility for refining GWAS loci. *The American Journal of Human Genetics, 100*(6), 954-959.

[14] Wu, Y., Zeng, J., Zhang, F., Zhu, Z., Qi, T., Zheng, Z., . . . Montgomery, G. W. (2018). Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nature communications, 9*(1), 918.

[15] Bennett, D. A., Buchman, A. S., Boyle, P. A., Barnes, L. L., Wilson, R. S., & Schneider, J. A. (2018). Religious orders study and rush memory and aging project. *Journal of Alzheimer's Disease*(Preprint), 1-28.

[16] Biological insights from 108 schizophrenia-associated genetic loci. (2014). *Nature, 511*(7510), 421-427. doi: 10.1038/nature13595

[17] Zuo, C., Shin, S., & Keleş, S. (2015). atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics, 31*(20), 3353-3355.

[18] Friedman, J., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw., 33*(1), 1-22.

[19] Kheradpour, P., & Kellis, M. (2013). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic acids research, 42*(5), 2976-2987.

[20] Ng, B., White, C. C., Klein, H. U., Sieberts, S. K., McCabe, C., Patrick, E., . . . De Jager, P. L. (2017). An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat Neurosci, 20*(10), 1418-1426. doi: 10.1038/nn.4632

[21] Ben-Shachar, D., & Karry, R. (2007). Sp1 expression is disrupted in schizophrenia; a possible mechanism for the abnormal expression of mitochondrial complex I genes, NDUFV1 and NDUFV2. *PLoS One, 2*(9), e817.