Statistical Analysis and Methods for Human -Omics Data

Statistical Analysis and Methods for Human –*Omics* Data

Yen-Chen Feng

A Dissertation Submitted to the Faculty of

The Harvard T.H. Chan School of Public Health

in Partial Fulfillment of the Requirements

for the Degree of *Doctor of Science*

in the Department of Epidemiology

Harvard University

Boston, Massachusetts.

May, 2017

Dissertation Advisor: Dr. Liming Liang Yen-Chen Feng

Statistical Analysis and Methods for Human –*Omics* Data

Abstract

Fast advancement in high-throughput technology has allowed screening of millions of molecular markers at multiple levels of the biological system in large samples to study the genetic basis and biological variation underlying complex traits and diseases. Such -omics data covers variation in the genome, epigenome, transcriptome, proteome, as well as metabolome. My dissertation projects take advantage of these rich sources of human multi-omics data, focusing on developing and applying statistical methods to answer questions that often arise in large-scale “-omic” epidemiology studies.

Single nucleotide polymorphisms (SNPs) are inherited genetic variations that may confer genetic predisposition towards complex diseases. Genome-wide association studies (GWAS) have been particularly successful in identifying numerous SNPs associated with non-Mendelian traits. GWAS of different traits also open up new opportunities to study the shared genetics across a range of phenotypes. In ***Chapter 1***, I will describe how we examined such relationship between Alzheimer’s disease (AD) and cancer using GWAS summary statistics and identified significant, positive genetic correlations of AD with specific cancer types.

Epigenetic modifications, including DNA methylation, are another crucial layer that regulates gene expression in a tissue-specific manner without changing the genetic

code. DNA methylation is involved in determining cell differentiation and is a marker of inhibited transcription. Studying cell-type specificity of DNA methylation in relation to diseases helps to identify the key cell type(s) for mechanistic follow-up. In ***Chapter 2***, I will describe a statistical method we developed to estimate cell-type-specific phenotype-methylation association when direct measurement of cell-specific methylation is not available, and the simulations and real data analysis we conducted to evaluate its performance.

Metabolome is a key endpoint linking genotype to phenotype that reflects perturbations from all levels of biological processes. Metabolomics data measured by the LC-MS experiments provides a powerful framework for studying disease mechanism and drug discovery, yet it often suffers substantial batch effect that makes cross-study comparison difficult. In ***Chapter 3***, I will illustrate an approach to normalizing metabolomics data across studies using the information from overlapping samples. We compared different normalization methods and identified quantile normalization as a preferred method to calibrate the cross-study deviation in metabolite distributions.

Table of Contents

Abstract .... **iii**

List of Figures .... **vii**

List of Tables .... **ix**

Acknowledgements .... **x**

CHAPTER 1. **1**

Investigating the genetic relationship between Alzheimer's disease and cancer using GWAS summary statistics

Introduction 2

Materials and Methods 4

Results 9

Discussion 18

Bibliography 23

Supplementary Materials 27

CHAPTER 2. **29**

Estimating cell-type-specific DNA methylation effects in the presence of cellular heterogeneity

Introduction 30

Method 31

Simulation 34

Real Data Application 40

Discussion..44

Bibliography 47

Supplementary Materials 49

CHAPTER 3. **51**

A strategy for cross-study normalization of metabolomics data with overlapping samples

Introduction 52

Materials and Methods...53

Results 57

Discussion..62

Bibliography 65

List of Figures

CHAPTER 1.

Figure 1.1 Genetic correlation between AD and each cancer type, estimated by cross-trait LD score regression 10

Figure 1.2 Annotation-specific genetic correlations (±SE) between AD and each cancer type 14

Figure 1.3 Relationship between SNP, gene expression, and observed phenotype(s) 16

CHAPTER 2.

Figure 2.1 Two-cell-type simulation results: estimation of cell-specific methylation effects when the true effects are β_1 = 1 and β_2 = 2 36

Figure 2.2 Two-cell-type simulation results: estimation of cell-specific methylation effects when the true effects are $\boldsymbol{\beta_1}$ = 0.01 and $\boldsymbol{\beta_2}$ = 1 37

Figure 2.3 Three-cell-type simulation results; $\boldsymbol{P}$'s = (0.15, 0.45, 0.4) 39

Figure 2.4 Distributions of the estimated methylation level for each cell type versus the measured methylation levels in purified cells at cg26787239 on the IL4 gene 43

CHAPTER 3.

Figure 3.1 Toy example: metabolite distributions before and after normalization among the overlapping samples 55

Figure 3.2 Distribution of selected known and matched unknown metabolites among the overlapping samples before normalization 58

Figure 3.3 Distribution of selected known metabolites among the overlapping samples after normalization comparing the four different methods..59

Figure 3.4 Change in MRE after normalization for the 83 known metabolites 59

Figure 3.5 Change in MRE and r^2 after normalization for the 638 matched unknown metabolites 60

Figure 3.6 Distribution of selected matched unknown metabolites among the overlapping samples after normalization 62

List of Tables

CHAPTER 1.

Table 1.1 Summary of Cancer (from GAME-ON) & AD (from IGAP) GWAS meta-analysis results data 7

Table 1.2 Co-heritability and genetic correlation between AD and each cancer type, estimated by cross-trait LD score regression 11

Table 1.3 SNPs with potential cross-phenotype effect on AD and two cancer types (overall breast and overall lung cancers) detected by Cross Phenotype Meta-Analysis (CPMA) 15

CHAPTER 2.

Table 2.1 Two-cell-type simulation results: point estimates and standard errors of cell-specific effects from the last iteration of the MCEM algorithm 35

CHAPTER 3.

Table 3.1 Cross-study normalization methods 57

Acknowledgments

I would like to first express my sincere appreciation to my advisor, Dr. Liming Liang, for his constant guidance, training, and kindness through the course of my doctoral studies. The bright thinking and extensive knowledge of Dr. Liang have inspired me in many ways over the years. I am also grateful to my committee members, Dr. Peter Kraft, for his insightful comments and suggestions that helped me tremendously to improve my research, and Dr. Frank Hu and Dr. Andrea Baccarelli, for their valuable discussion, assistance, and encouragement throughout this work.

I would like to thank all the collaborators of my projects, without whose effort and support this work would not be possible. It was through the communications and discussions with these different groups of researchers that helped enhance my experience and skills of conducting research and conveying ideas.

I am thankful to the members of the PGSG program for their continuous warmth and encouragement. This is the place that broadened my horizon into the fascinating field of molecular and statistical genetics and introduced me to a group of brilliant scientists from various backgrounds. I feel very lucky to be a part.

The journey of pursuing a doctorate is often challenging and sometimes lonely, and I could not have progressed down the path without the most supportive, inspiring, and genuine friends who are like my second family away from home. I thank and love them for all the laughter they have brought me over the past years and the confidence they have had in me when I lost mine.

Lastly, my forever and deepest gratitude is to my dear family, for their unconditional love, caring, and understanding that allows me to freely explore the world and move forward each step of the way with strength and positivity.

CHAPTER 1.

Investigating the genetic relationship between Alzheimer's disease and cancer using GWAS summary statistics

Abstract

Growing evidence from both epidemiology and basic science suggest an inverse association between Alzheimer's disease (AD) and cancer. We examined the genetic relationship between AD and various cancer types using summary statistics of genome-wide association studies (GWAS) from the IGAP and the GAME-ON consortia. Sample size ranged from 9,931 to 54,162; SNPs were imputed to the 1000 Genomes European panel. Our results showed a significant positive genetic correlation between AD and five cancers combined (colon, breast, prostate, ovarian, lung; $r_g = 0.17$, $P = 0.04$), and specifically with breast cancer (ER-negative and overall; $r_g = 0.21$ and 0.18, $P = 0.035$ and 0.034) and lung cancer (adenocarcinoma, squamous cell carcinoma and overall; $r_g = 0.31$, 0.38 and 0.30, $P = 0.029$, 0.016, and 0.006). Estimating the genetic correlation in specific functional categories revealed mixed positive and negative signals, notably stronger at annotations associated with increased enhancer activity. This suggests a role of gene expression regulators in the shared genetic etiology between AD and cancer, and that some shared genetic variants modulate disease risk concordantly while others have effects in opposite directions. This genetic overlap does not seem to be driven by a small number of major loci; no single SNP was found to have a cross-phenotype effect. Our study is the first to examine the co-heritability of AD and cancer leveraging large-scale GWAS results. The functional categories highlighted in this study

need further investigation to illustrate the details of the genetic sharing and to bridge between different levels of associations.

Introduction

Alzheimer's disease (AD) and cancer are complex diseases of aging that impose an enormous public health burden worldwide [1-3]. There is a growing understanding that these seemingly disparate conditions have substantial overlap. The pathophysiology of AD includes most if not all of the hallmarks of cancer, including abnormal cell cycle entry, metabolic deregulation, oxidative stress, DNA damage, inflammation, and angiogenesis [4]. All of these similarities suggest the diseases would be comorbid, but the weight of epidemiological evidence points to an unusual inverse association [5-10].

While it is difficult to know for sure that this "inverse comorbidity" represents a true association and is not the result of survival bias, there is convincing biological evidence for it. A transcriptomic meta-analyses using gene-expression data from relevant tissues found a substantial number of shared genes and their corresponding pathways to be upregulated in AD but downregulated in lung, colorectal and prostate cancers, and vice versa [11]. Differential expression of microRNAs between cancer and Alzheimer's disease has also been demonstrated [12]. A number of shared proteins and pathways have been identified that are differentially regulated by cancer cells and degenerating neurons. This includes the enzyme Pin, which is overexpressed in most cancers but depleted in AD [13]; tumor suppressor p53, which promotes apoptosis but protects against cancer [14]; and the Wnt cell survival pathway, which is activated in cancer but downregulated in AD [15]. Genetics play an important role in these underlying biological pathways, and therefore is expected to contribute to the inverse relationship between the two disorders either additively or through interaction with external factors [11, 16, 17].

However, beyond these three long-suspected but yet-to-be-confirmed candidates (Pin1, p53 and Wnt), very little is known about the genetic overlap between AD and cancer. Using genome-wide association study (GWAS) individual level data or summary statistics, one might be able to identify significant single nucleotide polymorphisms (SNPs) common to both disorders and estimate the cross-trait heritability. Existing methods based on genotype data, such as bivariate restricted maximum likelihood estimation (REML) as implemented in Genome-wide Complex Trait Analysis (GCTA) [18, 19] and genetic risk score profiling [20, 21], have been applied to a number of traits for estimating genetic correlations. Another approach is LD Score regression of summary statistics, as was recently applied to 24 traits to assess their pairwise genetic correlations [22]. Patterns of genetic overlap among 42 traits were also examined using a Bayesian approach [23]. No study has yet reported the genetic correlation between cancer and AD.

In the present study, we investigated the genetic overlap between AD and a variety of cancer types using SNP-trait GWAS summary statistics. We first estimated the genome-wide genetic correlation between the two diseases, then evaluated sharing heritability in specific functional categories, and finally tested cross-disease associations at individual SNPs. We used AD GWAS meta-analysis summary-level data acquired from the International Genomics of Alzheimer's Project (IGAP) and nine cancer GWAS meta-analysis results from the Genetic Associations and Mechanisms in Oncology (GAME-ON) consortium. There were 54,162 individuals included in the IGAP dataset and a sample size ranging from 9,931 to 33,832 among the GAME-ON datasets. All were imputed with over 7 million SNPs from the 1000 Genomes Project. This to our knowledge is the first study to investigate the genetic overlap between AD and specific cancer types using large-scale GWAS summary results where no individual genotype data is required.

Materials and Methods

Data: GWAS summary statistics for AD and each cancer type

Summary statistics of association analysis for late-onset AD were obtained from the International Genomics of Alzheimer's Project (IGAP; [24]). International Genomics of Alzheimer's Project (IGAP) is a large two-stage study based upon genome-wide association studies (GWAS) on individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data on 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyze four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's disease Initiative – EADI the Alzheimer Disease Genetics Consortium – ADGC The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium – CHARGE The Genetic and Environmental Risk in AD consortium – GERAD). European population reference haplotype data in the 1000 Genomes Project (2010 release) was used for genotype imputation, and genomic control correction was applied to each study before meta-analysis. In stage 2, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 & 2 (Table 1.1). Only stage1 data was used in the following analysis.

Summary statistics for cancers were acquired from the Genetic Associations and Mechanisms in Oncology (GAME-ON) consortium, which included meta-analysis results for nine cancer types (colon cancer, ER-negative breast cancer, overall breast cancer, aggressive prostate cancer, overall prostate cancer, ovarian cancer, lung adenocarcinoma, lung squamous cell carcinoma, and overall lung cancer (Table 1.1). Study designs included population-based, hospital-based, or family-based case-control studies. SNPs in individual studies were genotyped using Affymetrix or Illumina platform, and SNP imputation was performed using IMPUTE2, MiniMAC or MACH with 1000 Genomes Project (March 2012) data as reference. In each study, principle components were adjusted in association analysis to control for confounding by population stratification. Imputed SNPs were filtered according to imputation quality (r^2) before meta-analysis, which was implemented using METAL [25]. The final number of SNPs ranged from 9M to 15M across cancer types, the number of samples also varied to some extent, with colon cancer and ovarian cancers having a smaller number (~10K-13K) while prostate, breast, and lung cancers having a larger number of samples (~20K-30K).

Study subjects in IGAP and GAME-ON were all of European ancestry and originated from countries in Europe, Canada, the United States, or Australia. There is no sample sharing between AD and any of the cancer studies used in our analysis.

Estimation of genome-wide genetic correlation

Genetic correlation between AD and each cancer type was estimated by cross-trait LD Score regression [22]. This is a recently developed method based on GWAS summary statistics that quantifies the genetic covariance (ρ_g; analogous to co-heritability) between two traits by regressing the product of the z-scores ($z_{1j}z_{2j}$) from two studies of traits against the LD Score (l_j) for each SNP j, assuming both traits follow polygenic inheritance. Genetic correlations were obtained as $r_g = \rho_g \Big/ \sqrt{h_{g1}^2 h_{g2}^2}$ by normalizing genetic covariance by SNP-heritability h_{g1}^2, h_{g2}^2 for each trait estimated from single-trait LD Score regression [26]. AD and cancer as complex diseases likely possess a polygenic genetic architecture and therefore it is appropriate for using cross-trait LD score regression to estimate their genetic correlation. Empirical genetic correlations between AD and cancer were also calculated by taking Pearson's correlation coefficients of AD z-score and cancer z-score from all SNP to get an initial sense of the direction and magnitude of the genetic parameter and to be compared with the r_g estimates from LD Score regression.

The analysis was implemented using the LDSC v1.0.0 software package [26]. First, LD scores of all SNPs from individuals of European descent in the 1000 Genomes Project were computed. Next, genetic correlation of each cancer type with AD was estimated via cross-trait LD Score regression. Intercepts from cross-trait LD Score regression were constrained to zeros as there is no sample overlap, while single-trait intercepts were specified at their original values so as not to over-constrain residual confounding bias due to population stratification or other factors (e.g. cryptic relatedness).

Table 1.1. Summary of Cancer (from GAME-ON) & AD (from IGAP) GWAS meta-analysis results data

Dataset	#SNPs	#Study[1]	#Case[1]	#Control[1]	#Total samples	Imputation QC
GAME-ON GWAS						
All Colon	8,840,515	6	5,100	4,831	9,931	info ≥ 0.7; certainty≥0.9
Breast ER-negative	10,988,257	8	4,939	13,128	18,067	r^2 >0.3
Breast (overall)	11,099,926	11	15,748	18,084	33,832	
Prostate aggressive	9,671,146	6	4,450	12,724	17,174	r^2 >0.3
Prostate (overall)	9,760,825	6	14,160	12,724	26,884	
Ovarian (overall)	15,344,587	4	4,369	9,123	13,492	r^2 >0.25
Lung adenocarcinoma	8,897,683	6	3,718	15,871	19,589	r^2 ≥ 0.3 or info ≥ 0.4
Lung squamous cell carcinoma	8,909,656	6	3,422	16,015	19,437	
Lung (overall)	8,945,892	6	12,160	16,838	28,998	
IGAP GWAS						
AD (stage1)	7,055,881	4	17,008	37,154	54,162	r^2 ≥ 0.3 or info ≥ 0.3
AD (stage1&2)	11,632	15	25,580	48,466	74,046	

[1] Max. number; may differ by SNP

The number of overlapping SNPs between AD and each cancer dataset was around 5M to 6M. Before investigating the genome-wide relationship of AD with individual cancer types, we examined the genetic correlation between AD and "any cancer type" combined using five independent GAME-ON cancer data sets that do not share samples with one another, including that of colon cancer, overall breast cancer, overall prostate cancer, ovarian cancer, and overall lung cancer. The summary association statistics for "any cancer" were obtained by meta-analyzing GWAS summary statistics data from the five cancer types using METAL [25].

Estimation of annotation-specific genetic correlation

To characterize the genetic overlap at the level of functional categories, for each cancer type that showed significant genetic sharing with AD, we estimated genetic correlation between AD and cancer in eight large annotations using cross-trait LD score regression. These annotations included repressed region, introns, transcribed region, super enhancers, DNase I hypersensitivity sites (DHSs), and histone marks H3K27ac, H3K4me1, and H3K4me3 [27, 28]. Each of them contained more than 600,000 overlapping SNPs between the AD and cancer datasets that appropriates the use of LD score regression. For each annotation, we re-calculated LD scores for SNPs assigned to that particular category and then used the annotation-specific LD scores for estimating the AD-cancer genetic correlation.

Detection of individual SNPs associated with AD and cancers

Tests for cross-phenotype effects were carried out at individual loci to detect SNPs that show cross-phenotype (CP) associations with both AD and cancer, for the cancer types that have a significant genetic correlation with AD.

For each cancer type, among the SNPs overlapping between AD and cancer summary statistics, we started by picking out SNPs with a SNP-AD p-value < 0.001, then selecting SNPs every 100kb apart to mimic LD pruning and to appropriately evaluate statistical significance based on number of independent tests; SNPs selected within each window were those with the smallest SNP-AD p-values. Next, we looked for any additional signal from cancer beyond the existing SNP-AD association. The less stringent p-value cutoff was chosen to be consistent with the SNP filtering criteria in the

IGAP stage 2 meta-analysis. Bonferroni correction was used to correct for multiple testing.

To search for SNPs of a possible CP effect on AD and one or more cancer types, we also conducted individual SNP meta-analysis using Cross-Phenotype Meta-Analysis (CPMA; [29]) to explore if there is any SNP associated with some of the cancer types in addition to its correlation with AD. The filtered SNPs with a SNP-AD p-value < 0.001 again underwent distance pruning based on a window of 100kb. CPMA was performed among the remaining SNPs, followed by FDR control to correct for multiple testing. SNPs were assigned to genes via PLINK with SNP attributes--dbSNP build 129 and gene range list--hg19 for inference of a potential common biological process between the two traits. eQTL function for each top SNP was checked upon at the Genotype-Tissue Expression (GTEx) portal.

Results

Genetic correlation estimates between AD and cancer

We observed an overall positive genetic correlation of 0.172 between AD and the five cancers combined (colon, breast cancer, prostate, ovarian, and lung cancers; p-value = 0.04) estimated via cross-trait LD Score regression from the 6 million SNPs included in both GWASs (Table 1.2; Figure 1.1).

Stratifying by cancer type, ER-negative and overall breast cancer showed significant positive genetic correlations with AD at $r_g = 0.21$ (p-value = 0.04) and $r_g = 0.18$ (p-value = 0.03), respectively. Lung adenocarcinoma, lung squamous cell carcinoma, and overall lung cancer also had a prominent positive genetic correlation with AD at $r_g = 0.31$ (p-value = 0.03), 0.38 (p-value = 0.02), and 0.30 (p-value = 0.01). This implied that the two traits—AD and breast cancer, or AD and lung cancer—may share some common genetic background across the genome and the shared gene variants modulate the diseases risk in the same direction. On the contrary, the genetic correlation between AD and aggressive and overall prostate cancer were negative but not statistically significant (rg = -0.07 and -0.09; p-value = 0.54 and 0.20, respectively). The genetic correlation was around 0.1 between AD and all colon cancer and was slightly below zero between AD and ovarian cancer, and both estimates were not significant.

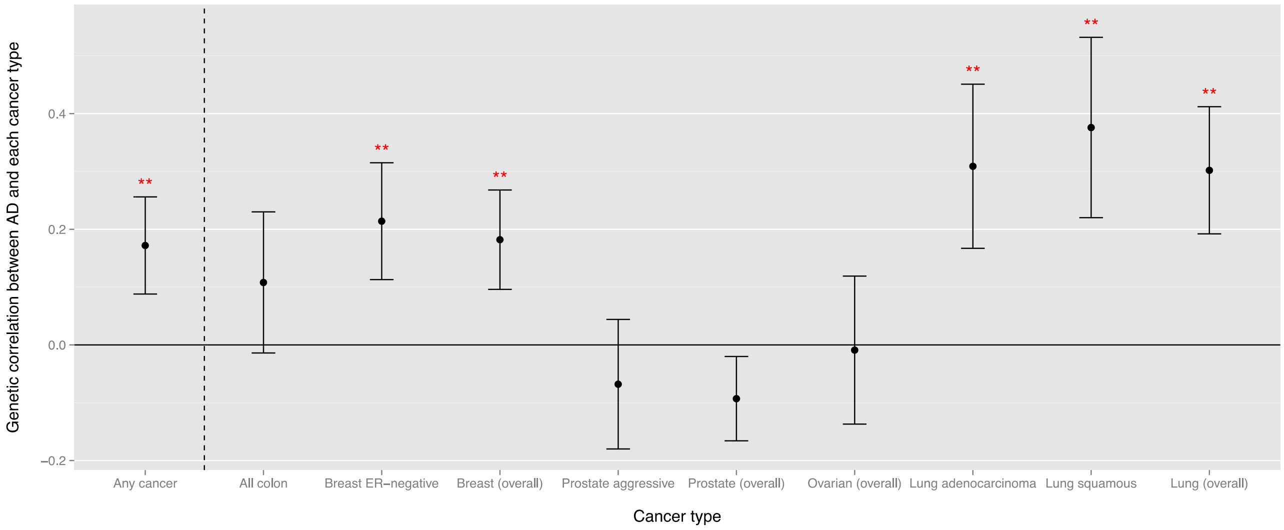


Figure 1.1. Genetic correlation between AD and each cancer type, estimated by cross-trait LD score regression

Error bars are displayed as point estimate ± SE; "**" denotes p-value for genetic correlation < 0.05; "Any cancer" category includes all colon cancer, breast cancer (overall), prostate cancer (overall), ovarian cancer (overall), and lung cancer (overall).

The genetic correlation estimates from cross-trait LD Score regression were consistent in terms of direction, relative magnitude, and statistical significance with our initial inspection of empirical correlation estimates between AD and each cancer type calculated as the Pearson's correlation coefficients between z-scores for all SNPs from the two traits (Table 1.S1), when LD between SNPs were not taken into account.

After learning the genome-wide relationship between AD and a variety of cancer types, we attempted to characterize the genetic sharing at regional and at individual SNP level between AD and the 2 cancer types that have a significant signal of genetic correlation, i.e. breast and lung cancers (overall).

Table 1.2. Co-heritability and genetic correlation between AD and each cancer type, estimated by cross-trait LD score regression[1]

Trait 1	Trait 2: Cancer	#SNPs[3]	co-h^2 (SE)	r_g	r_g SE	p-value
Alzheimer's Disease	Any cancer[2]	4,799,343	0.007 (0.003)	0.172	0.084	0.040
	All Colon	4,772,982	0.008 (0.009)	0.108	0.122	0.376
	Breast ER-negative	5,883,841	0.015 (0.007)	0.214	0.101	0.035
	Breast (overall)	4,743,056	0.012 (0.005)	0.182	0.086	0.034
	Prostate aggressive	5,405,868	-0.005 (0.008)	-0.068	0.112	0.543
	Prostate (overall)	5,666,977	-0.008 (0.006)	-0.093	0.073	0.204
	Ovarian	5,892,502	-0.005 (0.008)	-0.009	0.128	0.947
	Lung adenocarcinoma	5,681,123	0.015 (0.006)	0.309	0.142	0.029
	Lung squamous	5,681,315	0.018 (0.006)	0.376	0.156	0.016
	Lung (overall)	5,681,066	0.019 (0.005)	0.302	0.110	0.006

[1] All overlapping SNPs between AD-stage1 and cancer datasets were used. All cross-trait intercepts were constrained to 0, as there is no sample overlap.

[2] Any of the 5 overall cancer types: colon, breast, prostate, ovarian, and lung

[3] The number of overlapping SNPs between AD-stage1 and cancer datasets merged to the EUR LD score reference panel

Genetic correlation between AD and cancer by functional category

The first approach was evaluating the genetic correlation between AD and cancers by functional annotations to pin down specific regions on the genome that may explain more of the genetic sharing than other regions. This analysis additionally evaluated the annotation-specific relationship between AD and “any cancer type” where a notable positive r_g was also observed.

Our results showed that annotation-specific genetic correlations comprised of a mixture of positive and negative signals (Figure 1.2; Table 1.S2). Effect sizes of genetic variants in the repressed and the H3K4me3 annotations were negatively correlated between AD and breast cancer, lung cancer or the “any cancer” category, whereas positive genetic correlations were observed in the other six annotations. The only significant relationship appeared at super enhancers for that between AD and the five cancer types combined. Examining across all functional categories, three regions that represent active enhancer marks on the genome, including super enhancers, H3K27ac, and H3K4me1, all showed stronger and positive AD-cancer genetic correlations. This indicated a possible role of gene expression regulation with respect to enhancer activity in the shared genetic etiology between AD and cancer.

Cross-phenotype associations between AD and cancer

In order to investigate if cross-trait genetic correlation could be explained by major genetic loci, we went down to individual locus level to find pleiotropic SNPs associated with both AD and cancer, the existence of which may implicate common genetic pathways shared by the two diseases.

For each cancer type, we searched for any AD-related SNPs that also have an effect on cancer. A total of 11,788 out of the 4,743,056 SNPs common in both AD and breast cancer summary statistics and 14,655 out of the 5,681,315 AD-lung cancer overlapping SNPs remained after the filtering procedure (SNP-AD p-value < 0.001). There were 1507 SNPs present in both AD and breast cancer and 1648 SNPs present in both AD and lung cancer datasets after the 100kb window based pruning of SNPs. Among them, no SNP was significant after Bonferroni correction for breast cancer (top SNP rs59776273; chr4:47,792,047, SNP-breast cancer p-value = 9.9*10^{-5}). While for lung cancer there were two candidate SNPs that survived the correction: *rs56117933* at chr15:78,832,349 (unadjusted SNP-lung cancer p-value = 4.1*10^{-20}, corrected p-value = 6.7*10^{-17}), in close proximity to the *PSMA4* gene encoding for proteasome subunit alpha 4, whose polymorphisms have been related to lung cancer susceptibility from published GWAS [30], and *rs11249708* at chr5: 179821728 (unadjusted SNP-lung cancer p-value = 1.5*10^{-5}, corrected p-value = 0.025), for which no previous genome-wise associations have been reported.

We next carried out CPMA tests to find SNPs showing residual association with one or both cancer types, given its initial association with AD at SNP-AD p-value < 0.001. The results showed that, 11 out of 1458 SNPs after distance pruning had a CPMA p-value < 0.01, but only one of them passed the FDR 5% threshold (Table 1.3).

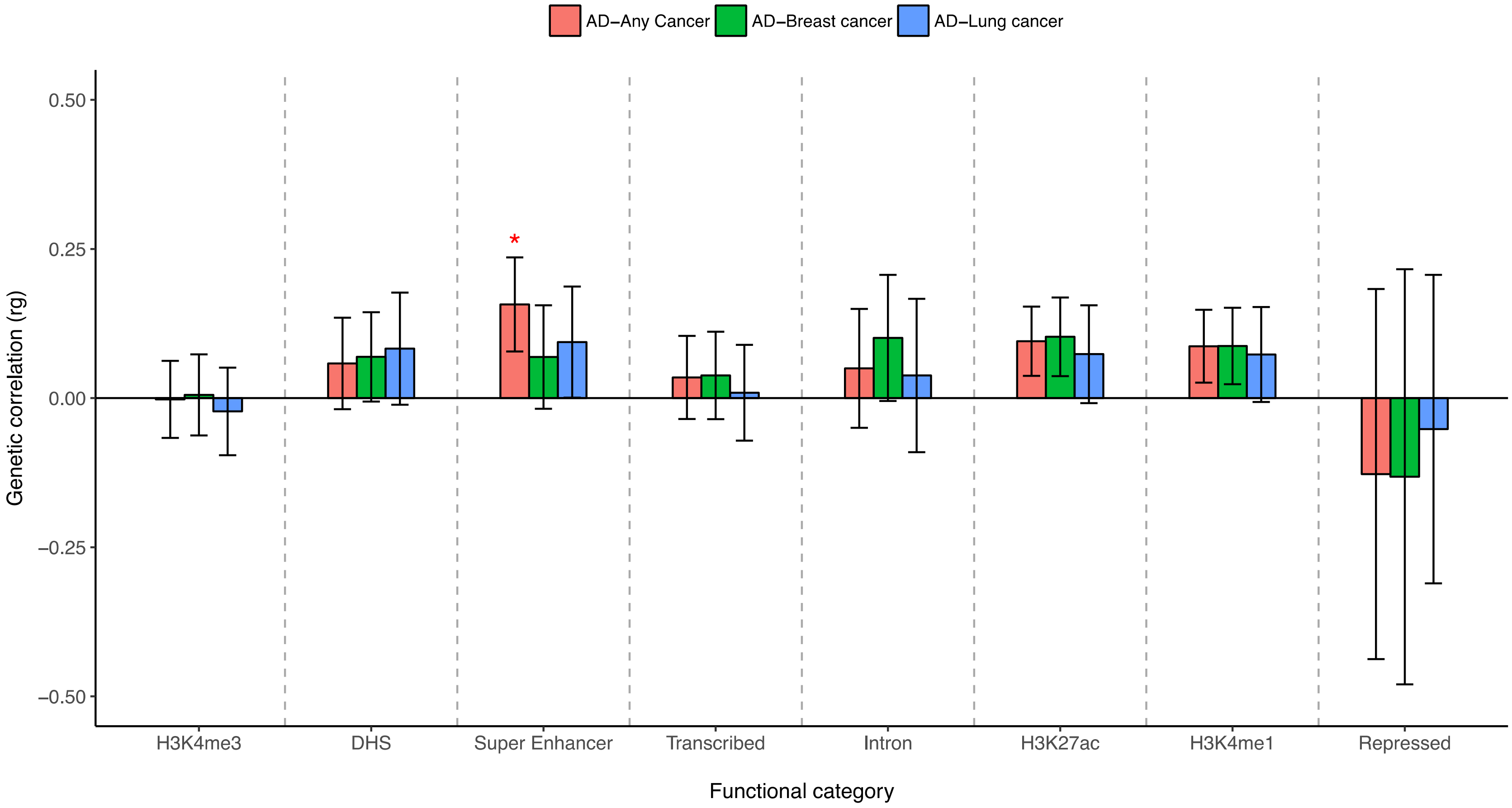


Figure 1.2. Annotation-specific genetic correlations (± SE) between AD and each cancer type "*" denotes p-value for genetic correlation < 0.05; functional categories on the x-axis were ordered based on its size, from the smallest (left) to the largest (right)

Table 1.3. SNPs with potential cross-phenotype effect on AD and two cancer types (overall breast and overall lung cancers) detected by Cross Phenotype Meta-Analysis (CPMA)

SNP	CHR	Position	Eff allele	Ref allele	AD		Breast cancer (overall)		Lung cancer (overall)		CPMA stat	CPMA p-value	FDR	Gene	Nearest gene
					z-score	p-value	z-score	p-value	z-score	p-value					
rs56117933	15	78832349	C	T	-3.34	8.3E-04	-0.59	5.6E-01	9.19	4.1E-20	73.98	<2.2E-16	<1.0E-15	-	PSMA4
rs11249708	5	179821728	A	G	-3.43	6.0E-04	1.53	1.3E-01	4.33	1.5E-05	14.79	1.2E-04	0.087	-	GFPT2
rs59776273	4	47792297	T	C	-3.52	4.4E-04	-3.89	9.9E-05	-2.14	3.2E-02	13.92	1.9E-04	0.093	CORIN (intron)	
rs17466060	8	27422740	A	G	4.60	4.3E-06	-2.39	1.7E-02	-3.45	5.7E-04	12.12	5.0E-04	0.182	-	EPHX2
rs3843702	15	80639403	A	G	-3.33	8.7E-04	-0.78	4.4E-01	3.82	1.3E-04	9.16	2.5E-03	0.575	-	LINC00927
rs3204635	12	57637593	A	G	-3.33	8.6E-04	-3.80	1.5E-04	-0.82	4.1E-01	9.10	2.5E-03	0.575	STAC3 (exon)	
rs7725218	5	1282414	A	G	-3.35	8.1E-04	-0.10	9.2E-01	3.97	7.2E-05	8.96	2.8E-03	0.575	TERT (intron)	
rs10896445	11	68967641	T	C	3.61	3.1E-04	-2.69	7.1E-03	2.55	1.1E-02	8.71	3.2E-03	0.577	-	MYEOV
rs77597338	2	53267773	G	A	4.39	1.1E-05	2.24	2.5E-02	2.84	4.5E-03	8.13	4.4E-03	0.705	-	ASB3
rs74766959	11	72065209	G	A	3.60	3.2E-04	1.39	1.7E-01	3.35	8.0E-04	7.88	5.0E-03	0.729	CLPB (intron)	
rs1568485	1	151736876	C	T	-3.30	9.7E-04	-3.53	4.2E-04	0.89	3.7E-01	7.63	5.7E-03	0.762	OAZ3 (intron)	

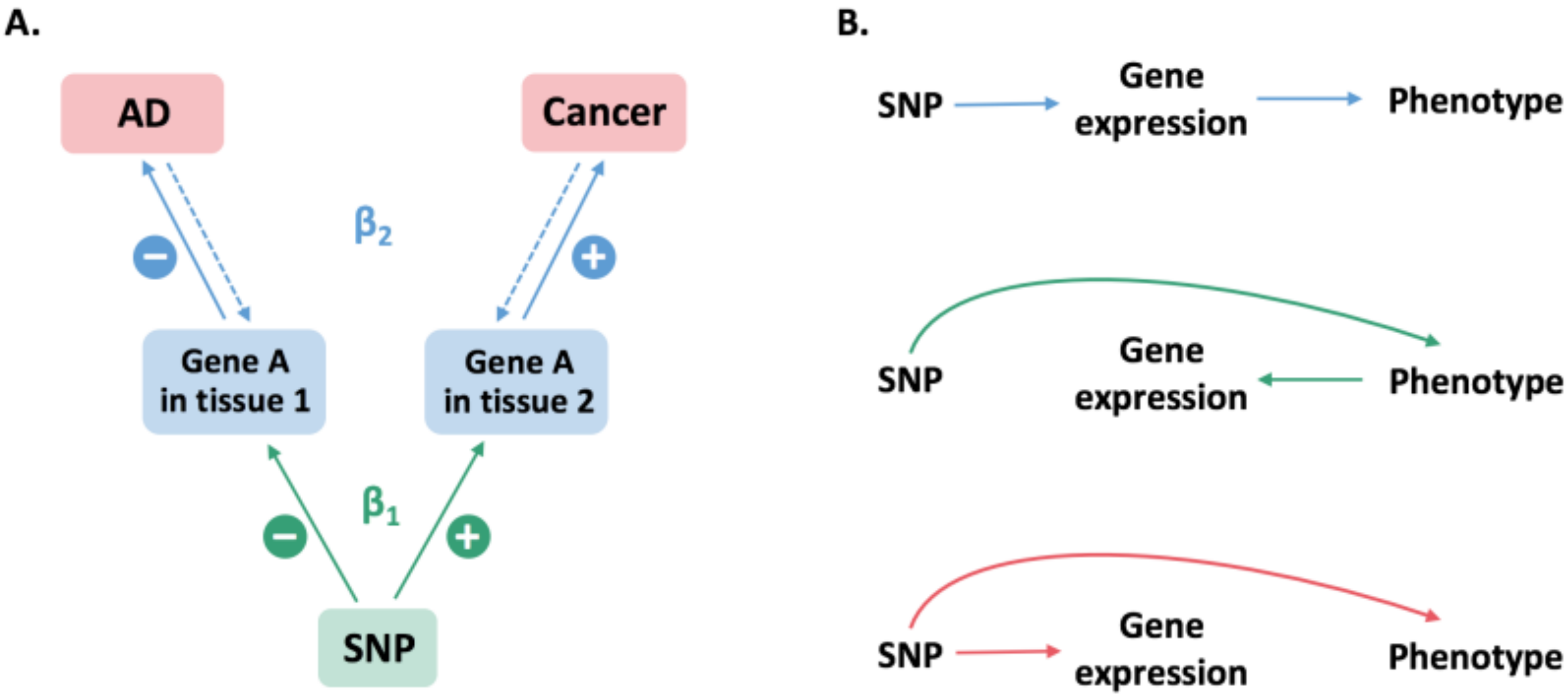


Figure 1.3. Relationship between SNP, gene expression, and observed phenotype(s)

(A) A possible scenario where an inverse correlation of gene expression effects [11] and a positive correlation of SNP effects between AD and cancer can be observed

(B) Possible causal pathways for the relationship between the three components if correlation exists between either two components. From top to down: causal effect of SNP on phenotype mediated through gene expression; gene expression reacts to phenotypic change due to SNP effect; pleiotropic effect of SNP on both gene expression and phenotype

This top SNP *rs56117933* (CPMA p-value < $2.2*10^{-16}$) was the same as discovered just previously, which had a highly significant association with lung cancer (p-value = $4.1*10^{-20}$) but a much larger p-value with breast cancer (>0.05). The significant AD SNP showing additional association with cancers was likely driven by one cancer type, similarly for the other 10 SNPs. This might reflect the heterogeneous nature of different cancer types and suggested to look for CP effects on AD and cancer independently by cancer type. The results also showed that cross-trait genetic relationships observed at the genome-wide level was not likely explained by several major variants, consistent with the polygenic architecture. Significant positive genetic correlations were found for between AD-breast cancer and AD-lung cancer, but as expected the SNP-AD z-score and the SNP-cancer z-score were not necessarily in the same direction. For example, SNP *rs17466060* appeared to increase AD risk (z-score = 4.60) but decrease the risk of both breast cancer (z-score = -2.39) and lung cancer (z-score = -3.45). No significant extols were found for the 11 SNPs in the most relevant tissues (brain or tumor) from the GTEx project, nor did they correspond to genetic variants on the previously reported candidate genes encoding p53, Pin1, or those involving in the Wnt pathway. The CP results on individual SNPs suggest that it would need a much larger sample size to obtain the same power as the cross-trait heritability estimate which aggregated information from all available SNPs on the genome or a particular functional category, and allow us to study the sharing genetic architecture of two diseases.

Discussion

In this study using data from two large GWAS consortia, we found a significant positive genetic correlation between AD and cancer overall, and specifically with breast and lung cancer. We also observed a suspected negative genetic correlation between AD and prostate cancer. These results establish that there is shared genetic variation between AD and cancer, but suggests that the direction of the genome-wide association may differ by cancer type. Examining the genetic correlation between AD and each cancer type in specific functional categories revealed that annotations linked to enhancer activity could play a role in the genetic sharing between the two diseases. These annotations may harbor important genetic variants involved in relevant pathophysiological pathways common to both AD and cancer. However, we did not identify any individual SNPs that had significant cross-phenotype associations with both diseases.

As we went from genome-wide investigation to a more local analysis of genetic relationship, we observed mixed signals of positive and negative directions of shared genetic effect within specific annotations. We also noted a discordance in effect size and direction across AD and cancers at the level of individual SNPs. This is expected, and confirmed that the overall aggregated genetic correlation is a sum of positive and negative genetic correlations due to different functional regions or individual variants. The power to detect shared genetic architecture at whole-genome, whole-functional category would be dependent on the consistency of effect direction of genetic variants in those categories or even the whole genome.

Our study found overall positive genetic correlations between AD and breast cancer and lung cancer, while epidemiological studies [5-7] and a transcriptome meta-analysis [11] suggest an overall inverse association of AD with many cancer types. This might be due to the fact that phenotypic comorbidity, correlation of expression effect and correlation of genetic effect are different levels of association that should not be expected to be the same. The inverse comorbidity of two diseases could be due to the joint effect of genetics and environment, where the non-genetic effect could be negatively correlated and have a larger magnitude than the positively correlated genetic effect. A possible scenario in which a negative AD-cancer association based on differential gene expression in relevant tissues [11] can co-exist with a positive genetic correlation among SNPs is depicted in Figure 1.3A; we note that this is only one of the numerous possibilities. In this case, the risk allele of a given SNP is associated with a decrease in expression of gene A in tissue 1 (egg. brain tissue) but an increase in its expression in tissue 2 (e.g. tumor tissue). An increased expression of gene A in tissue 1 is associated with a reduced risk of AD, while its higher expression in tissue 2 is associated with an elevated risk of cancer, resulting in inverse molecular comorbidity. This level of association can in fact be bi-directional. The net SNP effects on the two diseases would be positive ($\beta_{SNP} = \beta_1\beta_2$), and lead to a positive r_{ag} if the same is true for many more SNPs. In the analysis of partitioned co-heritability by functional categories, we observed both positive and negative genetic correlation in different categories. The functional annotation related to the negatively correlated category might explain the negatively correlated expression-AD association reported in previous studies and warrant further functional experiment.

Given the significant genome-wide associations of AD with some cancer types we have identified, we would need to gather gene expression data from brain and tumor tissues to establish a causal relationship linking SNP, gene expression, and both phenotypes together. Some possible scenarios for this are shown in Figure 1.3B. This would ideally be accomplished in eQTL studies that can evaluate which SNPs have a direct effect on the phenotypes, which SNPs have an indirect effect mediated by gene expression, how those SNPs affect or regulate gene expression levels to exert their influences on the phenotypes, and what genes are being regulated. eQTLs might also have different effect directions in tissues relevant to AD and tissues relevant to cancers. Integration of these results with other –omics data (e.g. Methylation QTL) would help to better understand the underlying molecular mechanism of shared genetics and how that could lead to the suggested AD-cancer comorbidity, thereby allowing definition of a more accurate link between the phenotypic association and the genetic association of AD with cancer.

In addition, we noticed that most of our genetic correlation estimates were of small magnitude and have a relatively large standard error (SE). This is likely due to sample size and from using summary level GWAS data instead of genotype data. It has been shown that genetic correlation in bivariate analysis (r_g) as a genetic parameter has a much larger sampling variance compared to proportion of phenotypic variance explained (h_g^2) by all SNPs in univariate analysis, which is true for both individual genotype data and in a pedigree design [31]. Simulation also showed that, when analyzing two case-controls studies of independent samples with an equal $h_g^2 = 0.2$ using genotype data, the power to detect an $r_g = 0.4$ with a sample size of 10,000 for each study is 0.9 and is only 0.4 when r_g = 0.2 [31]. Moreover, LD Score regression based on summary statistics generally yield bigger SEs than that from REML (GCTA) based on individual genotypes [22]. Together these suggest that an even larger sample size is required for LD Score regression as compared to REML (GCTA) to achieve comparable power when estimating r_g. The impact of sample size is evident in our results. We saw a larger SE around its r_g estimate for cancer subtypes of smaller number of cases (ER-negative breast cancer, aggressive prostate cancer, lung adenocarcinoma, and lung squamous cell carcinoma) relative to their overall cancer type counterparts (Table 1.1&1.2). The two cancer types that have the smallest sample size in our datasets—colon cancer GWAS with less than 10,000 and ovarian cancer GWAS with less than 15,000 individuals—were found to have a non-significant r_g surrounded by a wide confidence interval, but the effect size of r_g between colon cancer and AD is in fact not negligible. Increasing sample size would likely reduce SEs and increase statistical power to detect a true genetic correlation.

In conclusion, we identified significant genetic correlations between AD and certain types of cancer that indicate the presence of shared genetic variants and disease mechanisms between the two diseases. To the best of our knowledge, this is the first investigation of genome-wide association between AD and cancer using GWAS summary statistics coming from large scale studies. Our functional category analysis suggests that regulation of gene expression in relation to enhancer activity might play an important role in this shared heritability. Integration with gene expression data or eQTL studies in specific tissues is needed to better define the overlapping biological pathways, find genes and regions on the genome to be targeted for functional studies, and connect the missing dots from genetic comorbidity discovered using SNP data to the association observed at the levels of gene expression and phenotype. We anticipate incorporating our current findings of a quantified and characterized genetic relationship between AD and a range of cancer types into functional studies that can generate a better understanding of the pathophysiology of AD and cancer and provide insights into novel therapeutic possibilities for both diseases.

Bibliography

1. Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. Alzheimers Dement. 2007;3:186-191.
2. Thun MJ, DeLancey JO, Center MM, Jemal A, Ward EM. The global burden of cancer: priorities for prevention. Carcinogenesis. 2010;31:100-110.
3. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. Lancet. 2006;367:1747-1757.
4. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646-674.
5. Catala-Lopez F, Crespo-Facorro B, Vieta E, Valderas JM, Valencia A, Tabares-Seisdedos R: Alzheimer's disease and cancer: current epidemiological evidence for a mutual protection, Neuroepidemiology. 2014;42(2):121-2. doi: 10.1159/000355899. Epub 2014 Jan 3.
6. Driver JA, Beiser A, Au R, Kreger BE, Splansky GL, Kurth T, Kiel DP, Lu KP, Seshadri S, Wolf PA. Inverse association between cancer and Alzheimer's disease: results from the Framingham Heart Study. Bmj. 2012;344:e1442.
7. Musicco M, Adorni F, Di Santo S, Prinelli F, Pettenati C, Caltagirone C, Palmer K, Russo A. Inverse occurrence of cancer and Alzheimer disease: a population-based incidence study. Neurology. 2013;81:322-328.
8. Roe CM, Behrens MI, Xiong C, Miller JP, Morris JC. Alzheimer disease and cancer. Neurology. 2005;64:895-898.
9. Realmuto S, Cinturino A, Arnao V, Mazzola MA, Cupidi C, Aridon P, Ragonese P, Savettieri G, D'Amelio M. Tumor diagnosis preceding Alzheimer's disease onset: is there a link between cancer and Alzheimer's disease? J Alzheimers Dis. 2012;31:177-182.
10. Roe CM, Fitzpatrick AL, Xiong C, Sieh W, Kuller L, Miller JP, Williams MM, Kopan R, Behrens MI, Morris JC. Cancer linked to Alzheimer disease but not vascular dementia. Neurology. 2010;74:106-112.
11. Ibanez K, Boullosa C, Tabares-Seisdedos R, Baudot A, Valencia A. Molecular evidence for the inverse comorbidity between central nervous system disorders and cancers detected by transcriptomic meta-analyses. PLoS Genet. 2014;10:e1004173.
12. Holohan KN, Lahiri DK, Schneider BP, Foroud T, Saykin AJ. Functional microRNAs in Alzheimer's disease and cancer: differential regulation of common mechanisms and pathways. Front Genet. 2012;3:323.

13. Bao L, Kimzey A, Sauter G, Sowadski JM, Lu KP, Wang DG. Prevalent overexpression of prolyl isomerase Pin1 in human cancers. Am J Pathol. 2004;164:1727-1737.

14. van Heemst D, Mooijaart SP, Beekman M, Schreuder J, de Craen AJ, Brandt BW, Slagboom PE, Westendorp RG. Variation in the human TP53 gene affects old age survival and cancer mortality. Exp Gerontol. 2005;40:11-15.

15. Inestrosa NC, Toledo EM. The role of Wnt signaling in neuronal dysfunction in Alzheimer's Disease. Mol Neurodegener. 2008;3:9.

16. Demetrius LA, Simon DK. The inverse association of cancer and Alzheimer's: a bioenergetic mechanism. J R Soc Interface. 2013;10:20130006.

17. Tabares-Seisdedos R, Rubenstein JL. Inverse cancer comorbidity: a serendipitous opportunity to gain insight into CNS disorders. Nat Rev Neurosci. 2013;14:293-304.

18. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76-82.

19. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012;28:2540-2542.

20. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. The Lancet. 2013;381:1371-1379.

21. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460:748-752.

22. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, ReproGen C, Psychiatric Genomics C, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control C, Duncan L, Perry JR, Patterson N, Robinson EB, Daly MJ, Price AL, Neale BM. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47:1236-1241.

23. Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet. 2016;48:709-717.

24. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, DeStafano AL, Bis JC, Beecham GW, Grenier-Boley B, Russo G, Thorton-Wells TA, Jones N, Smith AV, Chouraki V, Thomas C, Ikram MA, Zelenika D, Vardarajan BN, Kamatani Y, Lin CF, Gerrish A, Schmidt H, Kunkle B, Dunstan ML, Ruiz A, Bihoreau MT, Choi SH, Reitz C, Pasquier F, Cruchaga C, Craig D, Amin N, Berr C, Lopez OL, De Jager PL, Deramecourt V, Johnston JA, Evans D, Lovestone S, Letenneur L, Moron FJ, Rubinsztein DC, Eiriksdottir G, Sleegers K, Goate AM, Fievet N, Huentelman MW, Gill M, Brown K, Kamboh MI, Keller L, Barberger-

Gateau P, McGuiness B, Larson EB, Green R, Myers AJ, Dufouil C, Todd S, Wallon D, Love S, Rogaeva E, Gallacher J, St George-Hyslop P, Clarimon J, Lleo A, Bayer A, Tsuang DW, Yu L, Tsolaki M, Bossu P, Spalletta G, Proitsi P, Collinge J, Sorbi S, Sanchez-Garcia F, Fox NC, Hardy J, Deniz Naranjo MC, Bosco P, Clarke R, Brayne C, Galimberti D, Mancuso M, Matthews F, Moebus S, Mecocci P, Del Zompo M, Maier W, Hampel H, Pilotto A, Bullido M, Panza F, Caffarra P, Nacmias B, Gilbert JR, Mayhaus M, Lannefelt L, Hakonarson H, Pichler S, Carrasquillo MM, Ingelsson M, Beekly D, Alvarez V, Zou F, Valladares O, Younkin SG, Coto E, Hamilton-Nelson KL, Gu W, Razquin C, Pastor P, Mateo I, Owen MJ, Faber KM, Jonsson PV, Combarros O, O'Donovan MC, Cantwell LB, Soininen H, Blacker D, Mead S, Mosley TH, Jr., Bennett DA, Harris TB, Fratiglioni L, Holmes C, de Bruijn RF, Passmore P, Montine TJ, Bettens K, Rotter JI, Brice A, Morgan K, Foroud TM, Kukull WA, Hannequin D, Powell JF, Nalls MA, Ritchie K, Lunetta KL, Kauwe JS, Boerwinkle E, Riemenschneider M, Boada M, Hiltuenen M, Martin ER, Schmidt R, Rujescu D, Wang LS, Dartigues JF, Mayeux R, Tzourio C, Hofman A, Nothen MM, Graff C, Psaty BM, Jones L, Haines JL, Holmans PA, Lathrop M, Pericak-Vance MA, Launer LJ, Farrer LA, van Duijn CM, Van Broeckhoven C, Moskvina V, Seshadri S, Williams J, Schellenberg GD, Amouyel P. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45:1452-1458.

25. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26:2190-2191.

26. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47:291-295.

27. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, Ripke S, Day FR, ReproGen C, Schizophrenia Working Group of the Psychiatric Genomics C, Consortium R, Purcell S, Stahl E, Lindstrom S, Perry JR, Okada Y, Raychaudhuri S, Daly MJ, Patterson N, Neale BM, Price AL. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47:1228-1235.

28. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjalmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B, Stahl E, Kahler AK, Hultman CM, Purcell SM, McCarroll SA, Daly M, Pasaniuc B, Sullivan PF, Neale BM, Wray NR, Raychaudhuri S, Price AL. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 2014;95:535-552.

29. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, Abecasis GR, Barrett JC, Behrens T, Cho J, De Jager PL, Elder JT, Graham RR, Gregersen P, Klareskog L, Siminovitch KA, van Heel DA, Wijmenga C, Worthington J, Todd JA,

Hafler DA, Rich SS, Daly MJ. Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet. 2011;7:e1002254.

30. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chen C, Goodman G, Field JK, Liloglou T, Xinarianos G, Cassidy A, McLaughlin J, Liu G, Narod S, Krokan HE, Skorpen F, Elvestad MB, Hveem K, Vatten L, Linseisen J, Clavel-Chapelon F, Vineis P, Bueno-de-Mesquita HB, Lund E, Martinez C, Bingham S, Rasmuson T, Hainaut P, Riboli E, Ahrens W, Benhamou S, Lagiou P, Trichopoulos D, Holcatova I, Merletti F, Kjaerheim K, Agudo A, Macfarlane G, Talamini R, Simonato L, Lowry R, Conway DI, Znaor A, Healy C, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature. 2008;452:633-637.

31. Visscher PM, Hemani G, Vinkhuyzen AA, Chen GB, Lee SH, Wray NR, Goddard ME, Yang J. Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. PLoS Genet. 2014;10:e1004269.

Supplementary Materials

Table 1.S1. Correlations of summary statistics (z-scores) between AD and each cancer type, with a block Jackknife p-value*

Trait 1	Trait 2: Cancer	Corr1 (of z-scores)	bjk p-value1
Alzheimer's disease	All Colon	0.002	0.583
	Breast ER-negative	0.008	0.069
	All Breast	0.009	0.089
	Prostate aggressive	-0.005	0.223
	All Prostate	-0.006	0.236
	Ovarian	-0.003	0.581
	Lung adenocarcinoma	0.009	0.025
	Lung squamous	0.012	0.003
	All Lung	0.015	0.001

*Adjacent SNPs on the same chromosome were divided into blocks; overall there were around 200 blocks across the genome with each block size of 25K-28K SNPs. Block jackknife estimates were obtained via a leave-one-block-out estimation procedure

Table 1.S2. Genetic correlation between AD and each significant cancer type in the eight functional categories

Annotation	Cancer type	Num_SNPs*	r_g	SE	P-value
Repressed	Any cancer	2,200,465	-0.127	0.310	0.681
	Breast	2,172,902	-0.132	0.348	0.705
	Lung	2,610,980	-0.052	0.259	0.841
H3K4me1	Any cancer	2,089,387	0.087	0.061	0.155
	Breast	2,062,737	0.087	0.064	0.173
	Lung	2,492,345	0.073	0.080	0.359
H3K27ac	Any cancer	1,879,775	0.095	0.058	0.100
	Breast	1,860,057	0.103	0.066	0.119
	Lung	2,232,313	0.074	0.082	0.369
Intron	Any cancer	1,907,385	0.050	0.100	0.616
	Breast	1,883,447	0.101	0.106	0.340
	Lung	2,254,094	0.038	0.129	0.768
Transcribed	Any cancer	1,698,138	0.035	0.070	0.619
	Breast	1,677,283	0.038	0.073	0.604
	Lung	2,003,018	0.009	0.080	0.910
Super enhancer	Any cancer	803,218	0.157	0.079	0.046
	Breast	794,399	0.069	0.087	0.427
	Lung	952,330	0.094	0.093	0.313
DHS	Any cancer	830,077	0.058	0.077	0.448
	Breast	818,119	0.069	0.075	0.356
	Lung	994,325	0.083	0.094	0.378
H3K4me3	Any cancer	631,807	-0.002	0.065	0.974
	Breast	625,431	0.005	0.068	0.936
	Lung	750,740	-0.022	0.074	0.762

*overlapping SNPs

CHAPTER 2.

Estimating cell-type-specific DNA methylation effects in the presence of cellular heterogeneity

Abstract

DNA methylation is an epigenetic modification that controls cell lineage and regulates gene expression. Signatures of DNA methylation differ across tissues and cell types, and cell composition can largely confound the association between phenotype and methylation when samples consist a mixture of cell populations (e.g. whole blood). Many statistical methods have been developed to adjust for this potential bias. More importantly, examining cell-type-specific DNA methylation effects can help identify the causal cell type(s) to follow up and gain insight into the underlying biology. However, purified cell types are usually not available in large scale studies due to impediment cost. In this work, we proposed a method to estimate cell-specific methylation-phenotype associations from unsorted whole tissue data where cell type proportions are also available. We used a framework that combines Monte Carlo EM algorithm and Metropolis-Hastings sampler to recreate the unobserved cell-specific methylation and to estimate its effect on phenotypes. Through simulations, we demonstrated that the method can successfully identify the true effects under various parameter settings, even when the causal cell type is rare. Application to a real dataset showed that cell-specific methylation pattern decomposed using the algorithm matches the directly measured methylation status in purified cell types. The method can be readily applied to existing EWAS datasets and is free of bias due to cell type distribution.

Introduction

DNA methylation is an important epigenetic modification that often acts to inhibit gene transcription by blocking the binding of transcription factors onto DNA [1]. Association between change in DNA methylation and phenotypes is therefore of interest to understand the underlying mechanisms leading from genetics to diseases or other traits [2, 3]. The pattern of DNA methylation varies largely across tissues and cell types, and so controls many of the cell-type-specific activities without changing the DNA sequence [4-6].

Collecting the most relevant tissue for a phenotype of interest would be ideal to study its association with DNA methylation profiles, but in reality it is very difficult to achieve, especially when sample size is large. Common sources of tissues in epigenome-wide association studies (EWAS) include peripheral blood, saliva, tumor...etc., that often consist of a heterogeneous collection of various cell types. Consequently, varying degrees of cell type proportions and cell-specific methylation levels among individuals could both pose an effect on the phenotype under study, and results from EWAS using cell mixture samples would face huge confounding by cell composition if it is not carefully accounted for [7-10].

Many methods have been developed to correct for the potential bias, using directly measured or estimated cell type proportions as a covariate in regression analysis [11-15]. However, very few have discussed estimating cell-specific methylation effects directly from a mixture of cells [16, 17]. Cell-type specific phenotype-methylation associations can help identify the “causal” cell type(s) for experimental follow-up to gain insight into the biological role of significant CpG loci. These cell-specific signals might be attenuated or masked when whole tissue methylation is used to make inference about differential methylation status [18]. Technology can sort out cell populations for methylation measurement, but at a very high cost. Therefore, we proposed a statistical method to estimate cell-specific effects which requires only whole tissue methylation data and information on cell composition.

This approach combines Monte Carlo Expectation-Maximization (EM) algorithm with Metropolis-Hastings sampler to reconstruct the "missing" cell-specific methylation status and to estimate their associations with phenotypes. We illustrated this method using simulations and then examined its performance on a real dataset where cell-specific associations have been reported.

Method

We addressed the proposed method in a simple scenario, assuming there are only two cell types in the cell mixture samples (e.g. whole blood). Let Y be the quantitative trait value of interest, Z the total DNA methylation level in whole blood at a CpG locus quantified in M-value, P the estimated cell type proportions, and X the *unobserved* cell-type-specific DNA methylation level. We used total methylation in M-value for estimation because it is more statistically tractable compared to another common metric, Beta-value, which measures the proportion of methylated molecules bounded between 0 and 1. M-value and Beta-value can be easily converted via M = $\log_2$[Beta/(1-Beta)] [19].

For each individual i $(i = 1, \ldots, n)$ at a given CpG site, X can be modeled as following a multivariate normal distribution $\boldsymbol{X} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, or

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{x_1}^2 & \rho\sigma_{x_2}\sigma_{x_1} \\ \rho\sigma_{x_2}\sigma_{x_1} & \sigma_{x_2}^2 \end{pmatrix}\right),$$

where μ and σ^2 are the mean value and variance of X, and ρ is the correlation between methylation levels in different cell types. The total methylation Z is simply a weighted average of cell-specific methylation levels, based on their proportions:

$$Z_i = P_{1i}X_{i1} + P_{2i}X_{i2} + \gamma_i, \;\; \gamma_i \sim N\left(0, \sigma_\gamma^2\right),$$

where $P_{1i} + P_{2i} = 1$. Assume X affects the trait through its marginal effects:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma_\epsilon^2),$$

and β's are the cell-specific effects we aim to estimate. The complete data likelihood is then the joint density of Y, Z, and X:

$$L(\theta|Y,Z,X) = f(Y,Z,X|\theta) = f(Y,Z|X,\theta)f(X|\theta) = f(Y|X,\theta)f(Z|X,\theta)f(X|\theta)$$

$$= \frac{1}{\sqrt{2\sigma_\epsilon^2\pi}} \exp\left(-\frac{\left(y-(\beta_1x_1+\beta_2x_2)\right)^2}{2\sigma_\epsilon^2}\right) \cdot \frac{1}{\sqrt{2\sigma_\gamma^2\pi}} \exp\left(-\frac{(z-(p_1x_1+p_2x_2))^2}{2\sigma_\gamma^2}\right) \cdot \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

where $\theta = (\sigma_\gamma^2, \sigma_\epsilon^2, \mu_1, \mu_2, \sigma_{x_1}^2, \sigma_{x_2}^2, \rho, \beta_1, \beta_2)$. This model can be easily extended to more than two cell types.

As deriving the conditional distribution $f(X|Y,Z,\theta)$, required for the E-step of the EM algorithm and from which X should be drawn, is not possible, we adopted the use of Metropolis-Hastings (M-H) algorithm to simulate the missing data X. Multiple Monte Carlo samples of X are drawn at each M-H step, which are then used together in the M-step to estimate θ. The Monte Caro EM algorithm (MCEM) works as follows:

(0) Initialization

Randomly initialize values for θ and X : $\beta^{(T=0)}, \mu^{(0)}, \Sigma^{(0)}$, ${\sigma_{\epsilon}^{2}}^{(0)}$, ${\sigma_{\gamma}^{2}}^{(0)}$, and $X^{(0)} \sim f(X|\theta) = MVN\left(\mu^{(0)}, \Sigma^{(0)}\right)$

(1) E-step (achieved by Monte Carlo simulation using M-H sampler)

At iteration T, run a Markov chain for each individual i:

- Generate a new value of X from its proposal function: $X_i^* \sim P(X|\theta) = MVN\left(X_i^t, \Sigma^{(T)}\right)$, where X^t is the current value
- Compute the acceptance probability based on the full joint density:

$$\alpha = \min\left(1, \frac{f(Y_i, Z_i, X_i^* | \theta^{(T)})}{f(Y_i, Z_i, X_i^t | \theta^{(T)})}\right)$$

- Accept the proposed X^* as X^{t+1} with probability α; operationally, $u \sim Unif(0,1)$,

$$\begin{cases} \text{if } u < \alpha \text{, accept the proposed value: } X_i^{t+1} = X_i^* \\ \text{if } u > \alpha \text{, reject the proposed value: } X_i^{t+1} = X_i^t \end{cases}$$

Discard burn-in values; run the chain until $j = 1, 2, \ldots, m^{(T)}$ samples of independent draws of $X_{i,j}^{(T)}$ are obtained. This procedure is ideally equivalently to

$$X_j^{(T)} \sim f\left(X \middle| Y, Z, \theta^{(T)}\right) \text{, where } j = 1, \ldots, m^{(T)}$$

(2) M-step

Compute a better estimate for θ by maximizing the complete data likelihood with respect to each parameter:

$$\hat{\theta}^{(T+1)} = \max_{\theta} \frac{1}{m^{(T)}} \sum_{j=1}^{m^{(T)}} \sum_{i=1}^{n} \log f\left(Y, Z, X_{i,j}^{(T)} \middle| \theta\right),$$

which is then used back in the E-step to update the values of X.

Repeat the E-step and M-step until convergence of $\hat{\theta}$ is observed.

Simulation

We performed simulations to examine the performance of the proposed method, assuming when there are two or three cell types in the sample.

Simulation: two cell types

In the two-cell-type scenario, we tested when methylation in the two cell types are differentially associated with trait Y $(\beta_1 = 1, \beta_2 = 2)$ or when the effect comes almost solely from one cell type $(\beta_1 = 0.01, \beta_2 = 1)$. P was generated from Beta distribution, with mean values of P_{1i} and P_{2i} varied to take the values (0.25, 0.75), (0.5, 0.5), or (0.75, 0.25). The true X was simulated from bivariate Normal distribution with μ_1 = 0.8, μ_2 = 1.3; $\sigma_{x_1}^2$= 0.2, $\sigma_{x_2}^2$ = 0.3; and ρ = 0.3. Y and Z were generated given the true values of $X, \beta,$ and P, under each of the β and P combinations, while σ_ϵ and σ_γ were fixed at 0.1 and 0.05. Sample size was $n = 500$ for all simulations. Here ρ was given and not estimated to evaluate how the decomposition into X behaves in a more controlled setting.

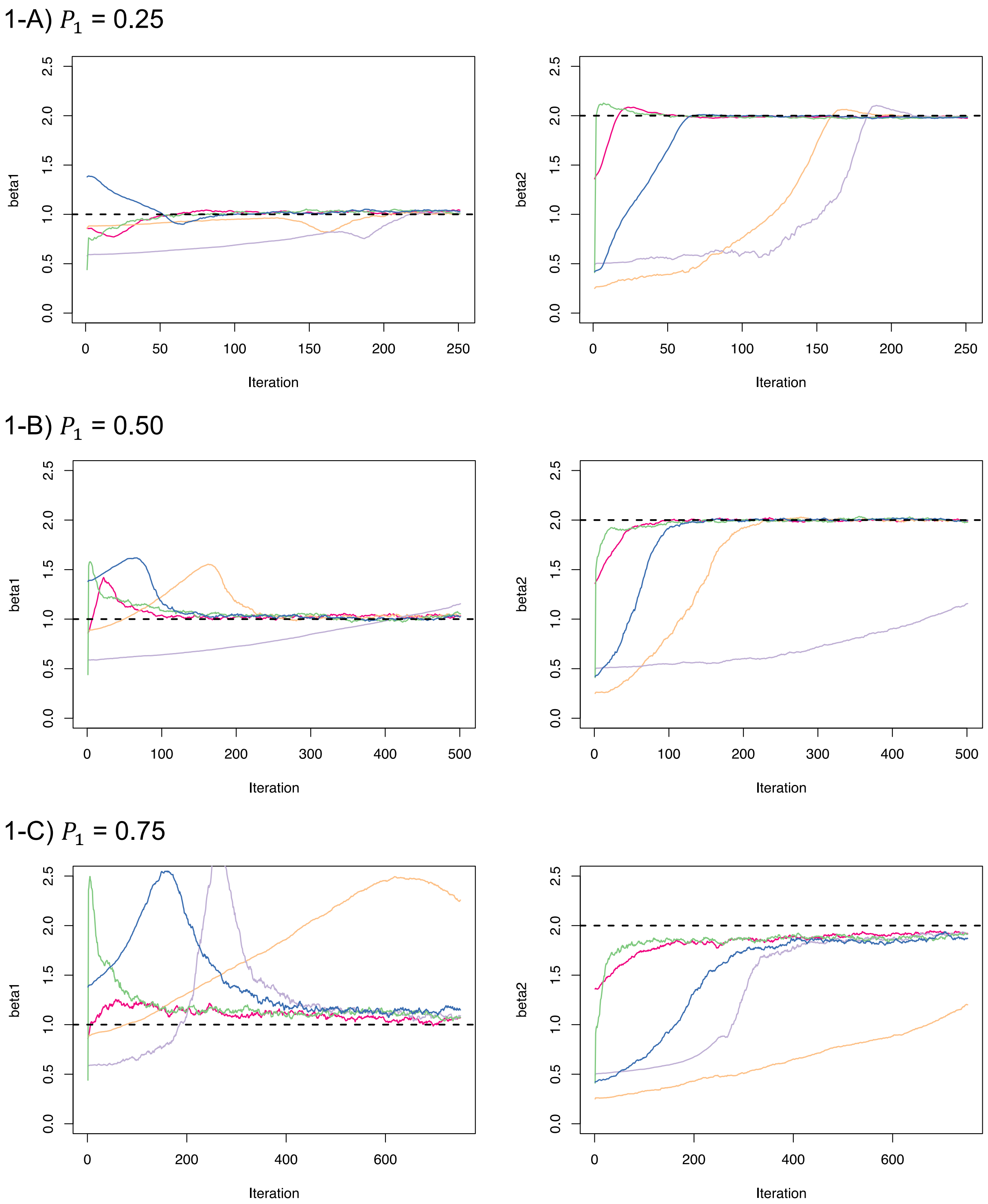
For each simulation setting, the MCEM algorithm was initialized with randomly selected values of θ and then $X \sim f(X|\theta)$; several different initial values were used to avoid finding only the local maxima. In the E-step at each iteration T, the first 100 burn-in values generated from the M-H sampler were discarded, and then each Monte Carlo sample was drawn every 100th values apart to minimize auto-correlation. Markov chain stopped once $m^{(T)} = 5$ samples of $X^{(T)}$ were obtained for estimation of $\hat{\theta}^{(T+1)}$ in the M-step. The procedure was repeated to iteratively estimate X and θ until convergence of parameter values. An average incomplete data likelihood was calculated over all $m^{(T)}$ samples at each iteration to help evaluate convergence of the parameters. Standard errors around the estimates were computed using 20 bootstrap samples by resampling the observed data (Y, Z, P). All the statistical procedures were conducted in R version 3.3.0.

Simulation results from the two-cell-type model showed an overall good performance of the proposed method. Cell-specific effects $\hat{\beta}$'s were correctly detected under all settings, even when a larger effect comes from the rare cell type (Figure 2.1&2.2). Given different initial values, most of the parameter estimates converged well to the true values, denoted by the dashed lines. In addition to $\hat{\beta}$'s, other parameters were also estimated with satisfactory accuracy (Figure 2.S1&2.S2). However, when the rare cell type has an effect size larger than that of the major cell type, time to convergence would be longer, and parameters would be estimated slightly less accurately and with more uncertainty (Figure 2.1&2.2; Table 2.1). Estimates that haven't converged or of less accuracy in general have a lower incomplete data likelihood compared to those closer to the true values (Figure 2.S2). Correlation between the reconstructed cell-specific methylation X at the last iteration and the true X of the same cell type was on average above 0.8 for all settings.

Table 2.1. Two-cell-type simulation results: point estimates and standard errors (SE) of cell-specific effects from the last iteration of the MCEM algorithm

Average cell type proportions	True $\beta_1 = 1, \beta_2 = 2$		True $\beta_1 = 0.01, \beta_2 = 1$	
	$\hat{\beta}_1$ (SE)[1]	$\hat{\beta}_2$ (SE)[1]	$\hat{\beta}_1$ (SE)[1]	$\hat{\beta}_2$ (SE)[1]
$P_1 = 0.25, P_2 = 0.75$	1.04 (0.04)	1.97 (0.03)	0.05 (0.04)	0.97 (0.03)
$P_1 = P_2 = 0.50$	1.03 (0.07)	1.99 (0.05)	0.06 (0.09)	0.99 (0.06)
$P_1 = 0.75, P_2 = 0.25$	1.07 (0.16)	1.92 (0.09)	0.07 (0.13)	0.90 (0.09)

[1] SEs were obtained via 20 bootstrap samples

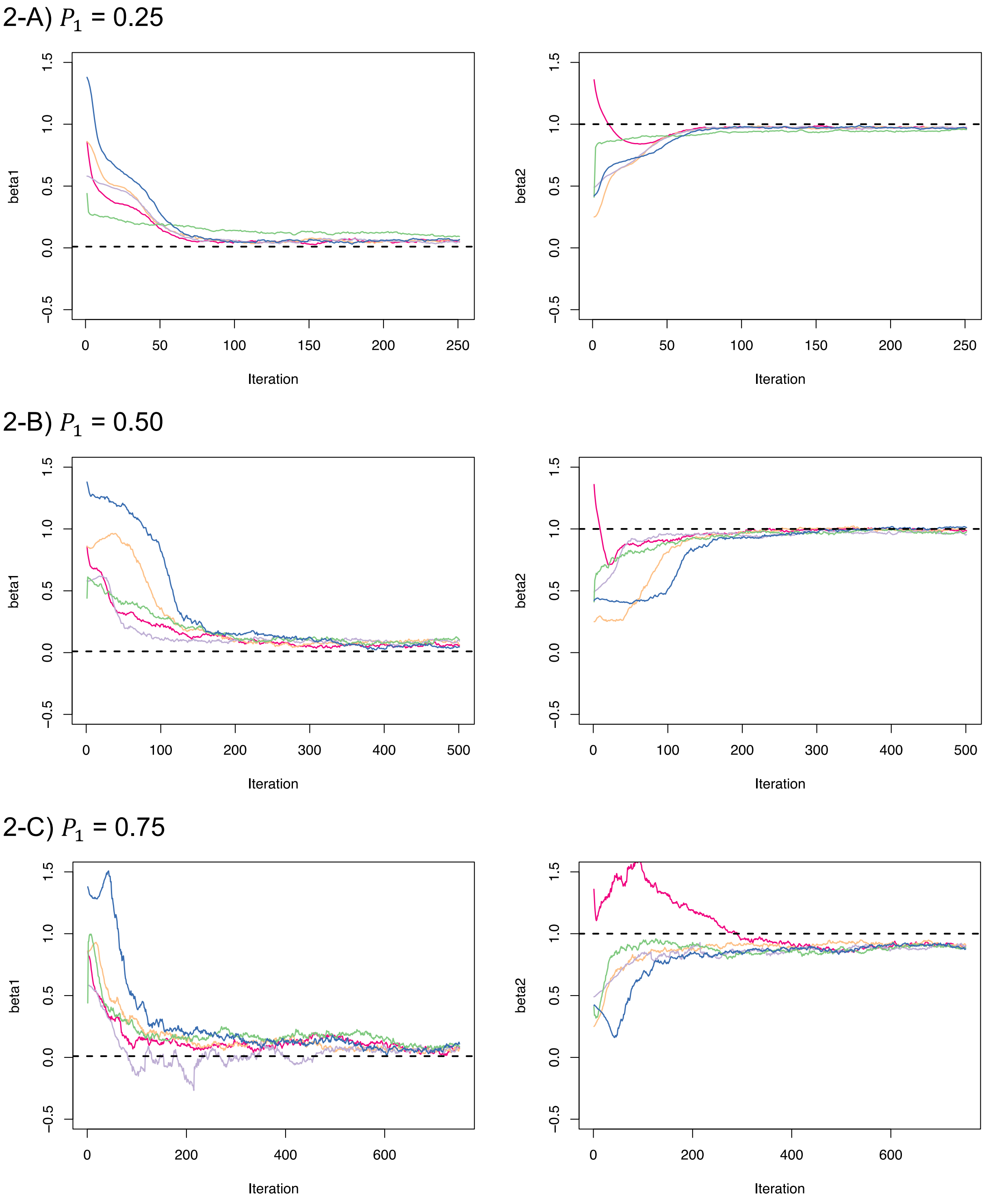


1-A) P_1 = 0.25

1-B) P_1 = 0.50

1-C) P_1 = 0.75

Figure 2.1. Two-cell-type simulation results: estimation of cell-specific methylation effects when the true effects are β_1 = 1 and β_2 = 2
Different colors indicate different initial values; dashed lines denote the true parameter values.



2-A) P_1 = 0.25

2-B) P_1 = 0.50

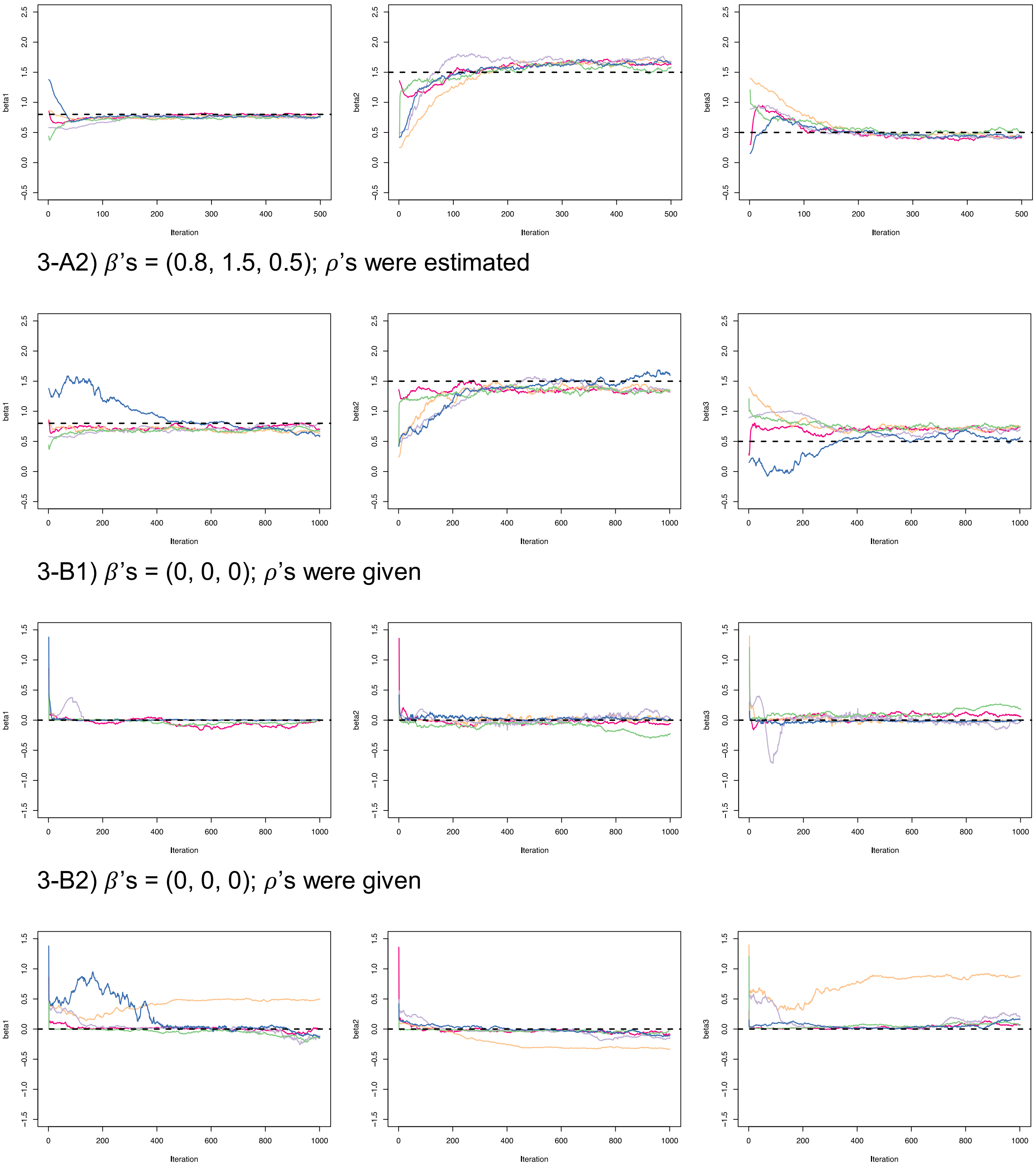
2-C) P_1 = 0.75

Figure 2.2. Two-cell-type simulation results: estimation of cell-specific methylation effects when the true effects are β_1 = 0.01 and β_2 = 1
Different colors indicate different initial values; dashed lines denote the true parameter values.

Simulation: three cell types

Extended to a three-cell-type model, we fixed the mean values of P_{1i}, P_{2i} and P_{3i} at (0.15, 0.45, 0.4), while varying cell-specific methylations to have different effect sizes on Y ($\beta_1 = 0.8, \beta_2 = 1.5, \beta_3 = 0.5$) or have no effect at all ($\beta_1 = \beta_2 = \beta_3 = 0$) when generating the data. The latter setting was to examine if parameters are identifiable even under a null model. The true values of X was simulated assuming that the rare cell type—with the smallest P—is a more active cell type with on average a lower methylation level and a larger variation (μ_1 = 0.5, μ_2 = 0.9, μ_3 = 1.0; $\sigma_{x_1}^2$= 0.25, $\sigma_{x_2}^2$ = 0.1, $\sigma_{x_2}^2$ = 0.15; and ρ_{12} = 0.19, ρ_{13} = 0.26, ρ_{23} = 0.57). σ_ϵ and σ_γ were again set to be 0.1 and 0.05 for generating Y and Z; sample size was 500. Parameter estimation was performed as described earlier. The correlation structure among cell-specific methylations was estimated empirically along with other parameters, and its results were compared to that when ρ's were given at its known values.

Under the three-cell-type scenario, the proposed algorithm was also able to arrive at an estimate close to the true value for $\hat{\beta}$'s and other parameters, both when all or when none of the cell-specific methylations is associated with Y (Figure 2.3). Results were overall comparable when correlations among cell-specific methylation (ρ's) were fixed versus when estimated directly. This suggests a fair use of empirical estimation of ρ when the true value is not known, with the caution that this might lead to more iterations required for parameter convergence and less stable and precise estimates.



3-A1) β’s = (0.8, 1.5, 0.5); ρ’s were given

3-A2) β’s = (0.8, 1.5, 0.5); ρ’s were estimated

3-B1) β’s = (0, 0, 0); ρ’s were given

3-B2) β’s = (0, 0, 0); ρ’s were given

Figure 2.3. Three-cell-type simulation results; P’s = (0.15, 0.45, 0.4)
Different colors indicate different initial values; dashed lines denote the true parameter values.

The promising results indicated the applicability of this method. We next applied it to a real dataset where cell-specific association was observed.

Real data application

The dataset we used is from Liang et al. [18], in which association of whole blood methylation with serum IgE level was found confined in eosinophils. The authors first identified and replicated an inverse association between IgE level and methylation in whole blood at 36 CpG loci. Further adjusting for cell composition revealed that IgE level was robustly associated with an increased number of eosinophils, but not with other cell types. Stratified by eosinophil cell counts and examining purified eosinophils showed study subjects with a higher IgE level had a lower level of methylation at these top CpGs and a greater number of eosinophils. This indicated an active role of eosinophils in the IgE and asthma pathophysiology. We aimed to verify the performance of the proposed approach by showing that eosinophils are a crucial cell population to follow up.

We combined whole blood methylation data from the discovery panel (MRCA) and one of the validation panels (SLSJ) for estimation, with a total sample size of 510. In each dataset, methylation status of each CpG was measured by Illumina HumanMethylation27 BeadChip in Beta-values of range 0 to 1. Measurements of cell counts of the five major cell populations in whole blood were available for all samples, including eosinophils (EOS), neutrophils (NEU), lymphocytes (LYM), monocytes (MON), and basophils (BAS). Mean values and standard deviations of the cell proportions were 4.6±3.9% (EOS), 53.1±9.9% (NEU), 34.1±8.3% (LYM), 7.5±2.0% (MON), and 0.6±0.7% (BAS).

To allow for meaningful interpretations of X and β, we adopted the following revised models to deconvolve the unsorted methylation data and to estimate the parameters: for a given CpG site,

1. Start with methylation status in Beta-values: $Z_i = \sum_k P_{ik} X_{ik} \sim Beta$, then the unobserved X_i's naturally take range from 0 to 1, where $i = 1, \ldots, n$ individuals and $k = 1, \ldots, K$ cell types.
2. Operate both X and Z on the M-value scale: $\boldsymbol{X}_i^* = M(\boldsymbol{X}_i) \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $Z_i^* = M(Z_i) \sim Normal$, where M = $\log_2$[Beta/(1-Beta)]. Z_i^* can be re-written as $Z_i^* = M(\sum_k P_{ik} X_{ik}) + \gamma_i = M(\sum_k P_{ik} M^{-1}(X_{ik}^*)) + \gamma_i$.
3. Evaluate the Y-X relationship via $Y_i = \sum_k \beta_k M^{-1}(X_{ik}^*) + \sum_{c=1}^{C} \beta_c C_{ic} + \epsilon_i$, where Y is the log-transformed IgE level, $c = 1, \ldots, C$ covariates, $M^{-1}(X_i^*)$ is the estimated cell-specific methylation in Beta-value, and $\hat{\beta}_k$ is the adjusted cell-specific methylation effect on Y.

The complete data likelihood was the joint density of (Y, Z^*, X^*). We excluded basophils from estimation for its rarity, and adjusted age and gender in the model ($K = 4, C = 2$). Pairwise correlations of the four cell-specific methylation levels were estimated empirically. We chose one of the top CpGs, *cg26787239* on the *IL4* gene, for demonstration. In the E-step, number of burn-in values was increased to 1000 to ensure values generated from the M-H sampler approximate those from the underlying conditional distribution of X, while all the other settings remained the same.

The results showed that, although likelihood space was bumpy for the estimates of cell-specific associations to converge well, the deconvolution algorithm led to stable and correct estimates of the cell-specific methylation status. Estimation started with different initial values all resulted in similar values; two of them were plotted in Figure 2.4A&B in which the final estimates had the largest likelihood. The estimated methylation levels (Beta-value) in the four cell types were directly comparable to the measured levels in purified cell populations reported in Renius *et al.* [5] and Liang *et al.* [18] (Extended Data Figure 2; Figure 2.4C). Both pointed to a lower methylation level with a wider variation for eosinophils at this IgE-associated CpG locus compared to other cell types, identifying eosinophils as an active cell type in relation to the etiology of IgE-associated asthma.

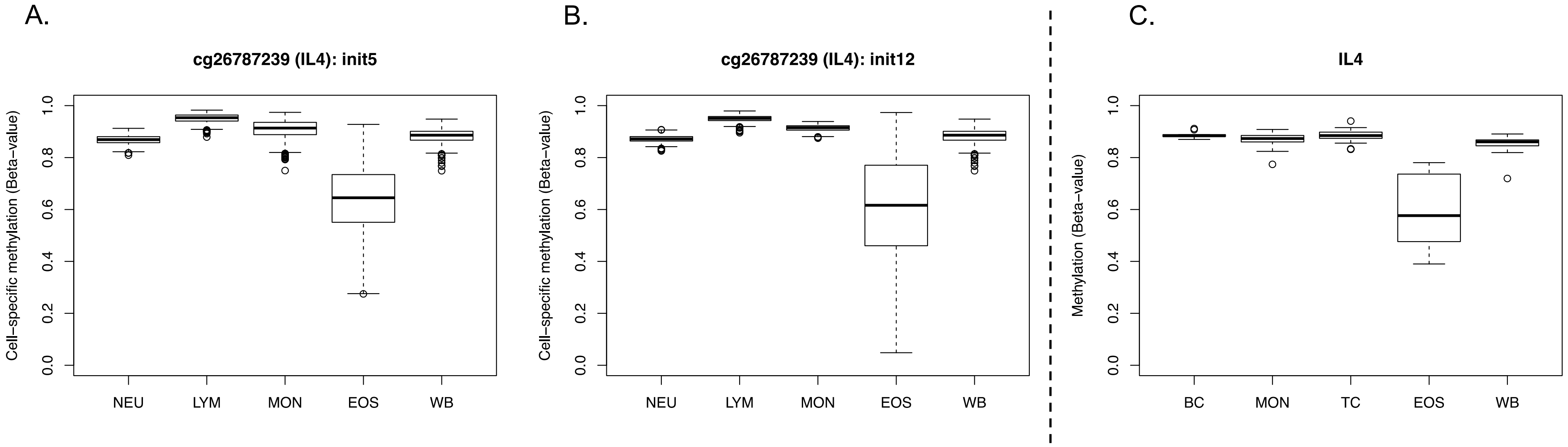


Figure 2.4. Distributions of the estimated methylation level for each cell type (A&B) versus the actual measured methylation levels in purified cells (C) at *cg26787239* on the *IL4* gene

(A&B) Two sets of the final estimates of X that led to the largest incomplete data likelihood, including the methylation status in neutrophils (NEU), lymphocytes (LYM), monocytes (MON), and eosinophils (EOS), in addition to the observed level in whole blood (WB)
(C) Directly measured methylations status in B cells (BC), monocytes (MON), T cells (TC), eosinophils (EOS), as well as in whole blood (WB)

Discussion

Detecting phenotype-methylation association in a cell-type specific manner provides insight into the underlying mechanism for CpGs identified from EWAS. In this work, we described a method to estimate methylation effects from each cell type in the presence of cellular heterogeneity, using an approach combining Monte Carlo EM algorithm and Metropolis-Hasting sampling. Through simulations, we demonstrated that the method can successfully identify the true effects and reconstruct the unobserved cell-specific methylation with good accuracy, even when the causal cell type is rare or when none of the cells has an effect. While application to a real dataset has not yielded solid estimates for the methylation effects, decomposing cell mixture methylation into cell-specific components revealed a consistent pattern with that of the actual observed methylation in purified cell types [5, 18].

Our method requires only measurements readily available in most of the EWAS, including phenotype information, methylation in unsorted, whole tissue samples (e.g. peripheral blood), and cell composition. Studies where direct measurement of cell counts or cell type proportions are not present, statistical methods, such as the Houseman algorithm, can be applied to efficiently quantify cell type distributions [14, 15]. This suggests a potentially wide applicability of our proposed approach onto existing datasets at no additional cost.

Nonetheless, we noted several caveats of the current models for improvement. First, the unsatisfactory convergence behavior of the cell-specific methylation effect estimates in real data analysis indicated parameters might be unidentifiable in practice, and the need of other model specifications to capture the latent data structure. An extreme example of unidentifiable effects would be that, when DNA methylation association only comes from one cell type, and if all the other cell types are highly correlated, then effect estimates for these other cell types might be a mixture of positive and negative values with the sum of them canceling out to be zero. The results would point to associations with multiple cell types when only one of them is causally related to Y. Possible extensions of the current $Y{\sim}X$ model can include to add in the cell type proportions P, whose associations with a range of phenotypes have been reported in EWAS [7, 10], or meaningful $X-P$ interaction terms.

Second, model complexity increases with the number of cell types, especially when effects from all cells are modeled at the same time. In such case, we can adopt a two-cell-type model, treating the target cell type as one category, pooling cell proportions from all the other cell types into a second group, to estimate effect for the cell of interest ($\hat{\beta}_1$). We can simply loop over all the possible cell types, changing the target cell group one at a time, to obtain effect estimates for each of them.

Third, standard errors (SE) of the parameters are not easy to obtain to evaluate statistical significance, because the conditional distribution $f(X|Y,Z,\theta)$ has no known closed form. Common methods used to estimate SE in the EM algorithm (e.g. Louis' method, SEM algorithm) relies on specifying the conditional distribution, while bootstrap resampling does not, it is computationally burdensome.

Lastly, considering time and efficiency, the proposed method is currently more suitable for candidate GpG analysis rather than epigenome-wide investigation. Time needed for obtaining reliable parameter estimates depends upon various factors, including number of cell types, length of the burn-in period, number of the Monte Carlo samples, whether the correlation structure is estimated or not, ...etc. In general, for a two-cell-type estimation that runs in parallel in R, it will take ~90min to complete every 250 iterations. We are planning to implement the algorithm in other programming languages (e.g. C++), which could potentially boost the speed for parameter and SE estimation and alleviate the computational load of the Monte Carlo simulation.

Despite these limitations, the method we described in this work provides a framework to estimate cell-specific DNA methylation effects that takes advantage of the existing datasets of whole tissue methylation and cell composition measurements. The MCEM estimation is conceptually straightforward, and the model specifications is flexible for extension. We plan to make the method more robust for application and the source code openly available for exploration.

Bibliography

1. Choy MK, Movassagh M, Goh HG, Bennett MR, Down TA, Foo RS. Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. BMC Genomics. 2010;11:519.
2. Baccarelli A, Bollati V. Epigenetics and environmental chemicals. Curr Opin Pediatr. 2009;21:243-251.
3. Bollati V, Baccarelli A. Environmental epigenetics. Heredity (Edinb). 2010;105:105-112.
4. Baron U, Turbachova I, Hellwag A, Eckhardt F, Berlin K, Hoffmuller U, Gardina P, Olek S. DNA methylation analysis as a tool for cell typing. Epigenetics. 2006;1:55-60.
5. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, Greco D, Soderhall C, Scheynius A, Kere J. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One. 2012;7:e41361.
6. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002;16:6-21.
7. Adalsteinsson BT, Gudnason H, Aspelund T, Harris TB, Launer LJ, Eiriksdottir G, Smith AV, Gudnason V. Heterogeneity in white blood cells has potential to confound DNA methylation measurements. PLoS One. 2012;7:e46705.
8. Houseman EA, Kelsey KT, Wiencke JK, Marsit CJ. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. BMC Bioinformatics. 2015;16:95.
9. Houseman EA, Kim S, Kelsey KT, Wiencke JK. DNA Methylation in Whole Blood: Uses and Challenges. Curr Environ Health Rep. 2015;2:145-154.
10. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 2014;15:R31.
11. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3:1724-1735.
12. Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics. 2011;27:1496-1505.
13. Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. Nat Methods. 2014;11:309-311.
14. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.

15. Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics. 2014;30:1431-1439.

16. Barfield R, Lin X: Estimating Cell-Type-Specific Associations from Whole Blood Methylation. in Joint Statistical Meetings2016.

17. Montano CM, Irizarry RA, Kaufmann WE, Talbot K, Gur RE, Feinberg AP, Taub MA. Measuring cell-type specific differential methylation in human brain tissue. Genome Biol. 2013;14:R94.

18. Liang L, Willis-Owen SA, Laprise C, Wong KC, Davies GA, Hudson TJ, Binia A, Hopkin JM, Yang IV, Grundberg E, Busche S, Hudson M, Ronnblom L, Pastinen TM, Schwartz DA, Lathrop GM, Moffatt MF, Cookson WO. An epigenome-wide association study of total serum immunoglobulin E concentration. Nature. 2015;520:670-674.

19. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.

Supplementary Materials

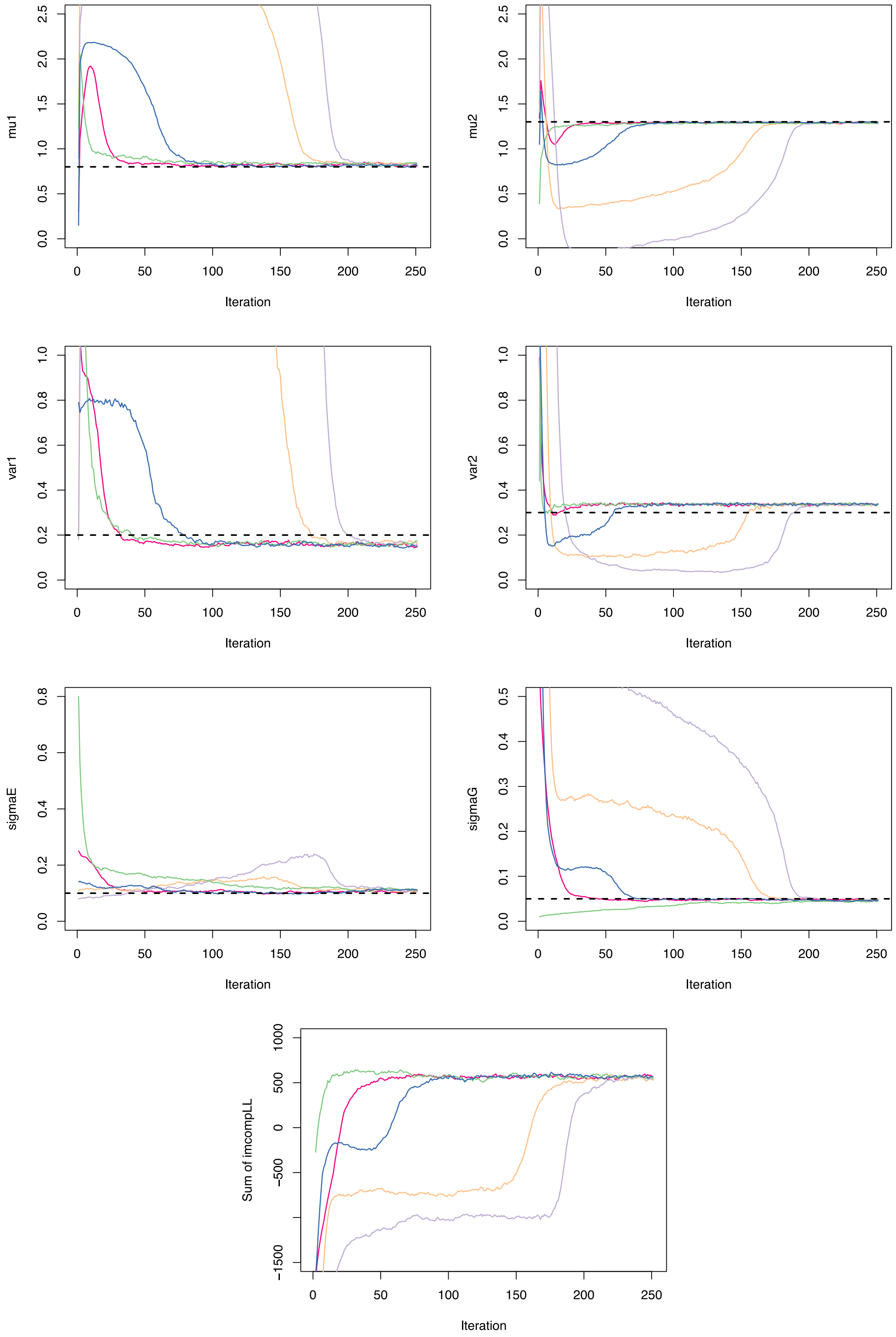


Figure 2.S1. Two-cell-type simulation results: estimation of other parameters and the incomplete data likelihood when the true effects are β_1 = 1 and β_2 = 2 at P_1 = 0.25 Different colors indicate different initial values; dashed lines denote the true parameter values.

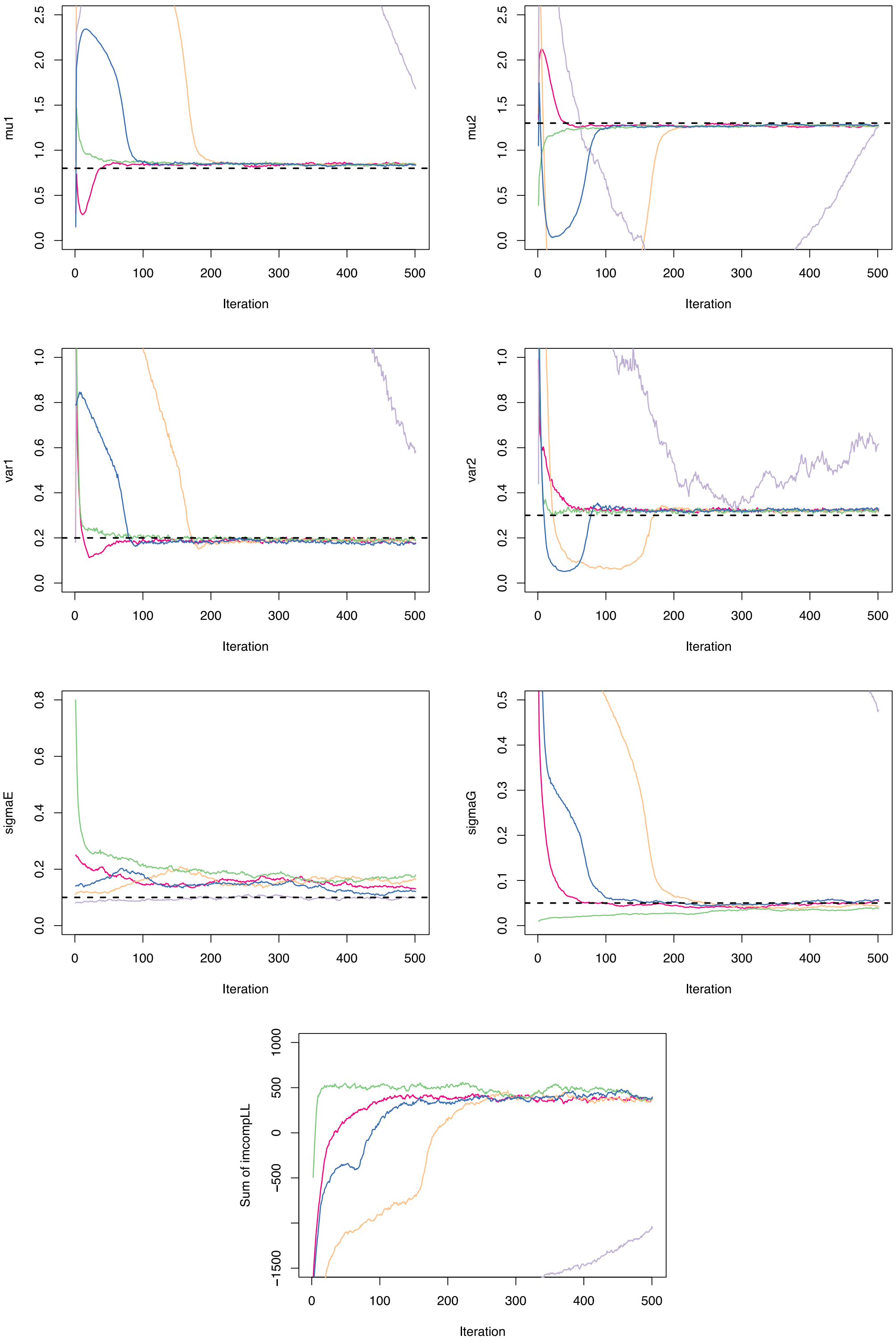


Figure 2.S2. Two-cell-type simulation results: estimation of other parameters and the incomplete data likelihood when the true effects are β_1 = 1 and β_2 = 2 at P_1 = 0.50 Different colors indicate different initial values; dashed lines denote the true parameter values.

CHAPTER 3.

A strategy for cross-study normalization of metabolomics data with overlapping samples

Abstract

Metabolomic profiling using liquid chromatography–mass spectrometry (LC-MS) provides a power tool to study how thousands of actionable metabolites in the biological system are related to disease mechanisms and gene functions. However, data generated from the LC-MS assay often suffers sizable batch effects and noises, making combining samples across studies difficult and warranting extensive preprocessing steps. We proposed to correct the cross-study difference in metabolite profiles due to technical variation using distributions of overlapping samples shared across studies, and examined the performance of median normalization, robust regression, and quantile normalization in calibrating two metabolomics data that have 260 identical biological samples. We showed that quantile normalization outperformed other methods in reducing the between-study batch effect, reflected by an overall decrease in mean relative error and increase in r^2 for both targeted and untargeted metabolites among the overlapping samples. This approach can benefit many existing pilot metabolomics studies where shared samples are available to increase sample size and power for detecting metabolite-phenotype associations.

Introduction

Metabolomics characterizes the complete profile of small-molecule metabolites that reflects changes from all levels of the biological system, as well as from environmental exposures [1-3]. Advancement in metabolomics technology, such as liquid chromatography–mass spectrometry (LC-MS), has allowed large-scale epidemiological studies to investigate the relationship between thousands of metabolites and disease mechanisms [3-7]. However, metabolite measurements as generated from the LC-MS experiments often face various sources of systematic biases, such as batch effect and time-dependent performance of the LC columns [8, 9], which warrants extensive data preprocessing before any analysis attempt, and renders combining samples cross studies a particular challenge.

Although software provided by commercial platforms has built-in normalization procedure for within-batch or within-study normalization, there is currently no optimal solution to the cross-study recalibration problem. We observed the use of shared samples across several metabolomics studies in epidemiological research, including biological samples or identical technical samples inserted to ensure comparable analytical performance [10], and recognized this as an opportunity for cross-study normalization. Given metabolites from the same sample should have the same distribution, the idea is to fix the metabolite distribution among the overlapping samples in one study as the reference, and to shift its distribution in another study toward the reference, making them overall comparable. Samples can be combined after calibration to increase statistical power for association detection.

We demonstrated the strategy using median normalization, robust regression, and quantile normalization on two LC-MS metabolomics data that share identical biological samples. We examined the pre- and post-normalization concordance among the overlapping samples for both targeted and untargeted metabolites to suggest the optimal method for calibrating metabolites across studies.

Materials and Methods

Motivating datasets

To compare the performance of different normalization methods, we used two polar metabolomics datasets generated for the PREDIMED trial [11-13]. These include two case-cohort studies nested within the trial, one focusing on cardiovascular disease (CVD) and the other examining type 2 diabetes (T2D) as the primary outcome [14, 15]. Metabolite profiles of study samples at baseline and at year 1 visit were measured by the LC-MS method in the positive ionization mode using both the targeted and untargeted platforms at the Broad institute. Pooled plasma samples of every 20 study samples were also measured as a reference for within-data calibration. The CVD study consists of 1,994 biological samples and 101 pooled plasma samples, and for each sample measurements of 83 targeted metabolites and 3,209 untargeted metabolites. The T2D metabolomics study contains measurements of the same 83 known metabolites as well as 7,197 untargeted metabolites from 1,963 study samples and 93 pooled plasma samples. Samples were collected at the same time, but metabolomics experiments for the two studies were conducted nearly two years apart. The two studies share 260 overlapping samples from 164 individuals.

Matching of untargeted metabolites across studies

To identify a common set of unknown metabolites for normalization, we first applied quality control on the untargeted metabolomics data for each study separately, and then matched the remaining metabolites of unknown identity across studies at the suggested threshold of mass-to-charge ratio (M/Z) < 0.0005 and retention time (RT) < 0.2. The exclusion criteria included: (1) missing rate > 10% among study samples, (2) missing rate > 20% among pooled plasma samples, (3) coefficient of variation(CV) > 25% among pooled plasma samples, and (4) identified by the MetProc [16] package as a likely measurement artifact. After QC, 1,132 metabolites remained in the CVD untargeted metabolomics data, and 3,172 untargeted metabolites were left in the T2D data. Any known metabolites were also removed. Matching on the filtered data resulted in 638 likely identical unknown metabolites between the two studies.

Methods to normalize metabolites across studies

We used the cross-study panels—metabolite measurements of the 83 known and the 638 unknown metabolites from the 260 overlapping samples in each study—to address the idea for cross-study normalization. For a given metabolite, suppose X is its distribution among the overlapping samples in study 1 and Y in study 2, we would expect a systematic shift between X and Y (Figure 3.1; top), even though they are in fact repeated measurements and should match exactly. To adjust for the between-study batch effect, we proposed to fix one panel as the reference, and shift the distribution in the other toward the reference, making them overall comparable (Figure 3.1, bottom). In other words, the difference between *X* and *Y* can be used to build a normalization factor, equation, or function for cross-study normalization. If the strategy works as desired among the overlapping samples, it can then be applied to normalize metabolite levels for other samples unique to each study.

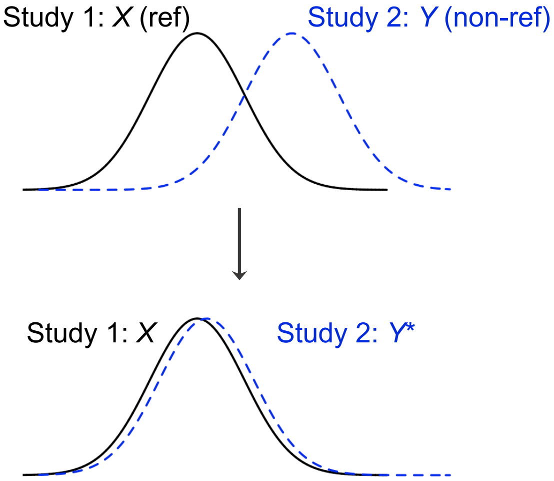


Figure 3.1. Toy example: metabolite distributions before (top) and after (bottom) normalization among the overlapping samples

Distribution in study/panel 1 (*X*) is used as the reference to normalize the distribution in study/panel 2 (original *Y* → normalized *Y**).

Methods used to realize the idea included median normalization, robust regression, and quantile normalization (Table 3.1). Median normalization takes the median of the differences in the two metabolite distributions as a normalization factor for repositioning the non-reference distribution, currently the most widely used approach; robust regression constructs a one-to-one normalization equation for metabolite values across panels while down-weighting the effects from outliers; and quantile normalization maps every value in the non-reference panel to the corresponding quantile of the empirical distribution in the reference panel to make the two distributions identical, a technique frequently used in microarray data.

We implemented the following strategy to examine and compare their performance in reducing the systematic deviation and improving the correspondence among overlapping samples:

(1) Divide the 260 samples into independent training and testing sets by an 1:1 ratio; samples from the same individual are always separated

For each metabolite,

(2) Use the training set distribution with less variation (smaller standard deviation) as the reference (X_{train}), and normalize metabolite values in the non-reference testing set (Y_{test}) using one of the three methods (Table 3.1)

For quantile normalization that involves the use of an empirical distribution function (*F*), we additionally tested when *F* from all samples is used (F_X, F_Y), as compared to that based on the overlapping samples.

(3) Calculate the pre- and post-normalization mean relative error (MRE) and r^2 for each method in the non-reference testing set (MRE = $\frac{1}{n_s}\sum_{i=1}^{n_s}\frac{|x_i-y_i|}{x_i}$, where *i* = 1, ..., n_s overlapping samples)

Repeat the process 200 times to obtain an average normalized value for each sample in the non-reference panel, as well as an average post-normalization MRE and r^2, to evaluate the overall performance. Method(s) that reduces MRE and increases r^2 more effectively after normalization would be a more ideal approach in calibrating metabolites across studies.

Table 3.1. Cross-study normalization methods (i = 1, ..., n_s overlapping samples)

Method	**Normalization function**	**Normalized value**
Median normalization	$M = \text{median}\left(y_{i,train} - x_{i,train}\right)$	$y_{i,test}^{*} = y_{i,test} - M$
Robust regression	$x_{i,train} = \mu + \beta * y_{i,train}$	$y_{i,test}^{*} = \hat{\mu} + \hat{\beta} * y_{i,test}$
Quantile normalization	$F_{X_{train}}, F_{Y_{train}}$	$y_{i,test}^{*} = F_{X_{train}}^{-1}\left(F_{Y_{train}}\left(y_{i,test}\right)\right)$
	F_X, F_Y	$y_{i,test}^{*} = F_X^{-1}\left(F_Y\left(y_{i,test}\right)\right)$

Results

Before normalization, metabolite distributions showed an overall good concordance among the overlapping samples for a majority of (~60) the known metabolites (Figure 3.2A), but a number of them (~10-15) were highly discordant (Figure 3.2B). Median r^2 for these 83 known metabolites across shared samples was 0.7 and median intraclass correlation (ICC) was 0.8. We confirmed the systematic difference was not due to true associations with the phenotype(s) but most likely detection limit of the LC-MS experiments. For unknown metabolites, half of them had an $r^2 > 0.5$ (Figure 3.2C), while many were very poorly aligned (Figure 3.2D), with a median ICC < 0.1. Only 134 of the 638 metabolites had an ICC > 0.4. The expected between-study difference confirmed the need for cross-study normalization to facilitate comparison and to combine samples together. The larger inconsistency and lower reliability of the matched unknown metabolites across panels also suggested that some of the matched pairs might not be the same molecule but merely measurement noise.

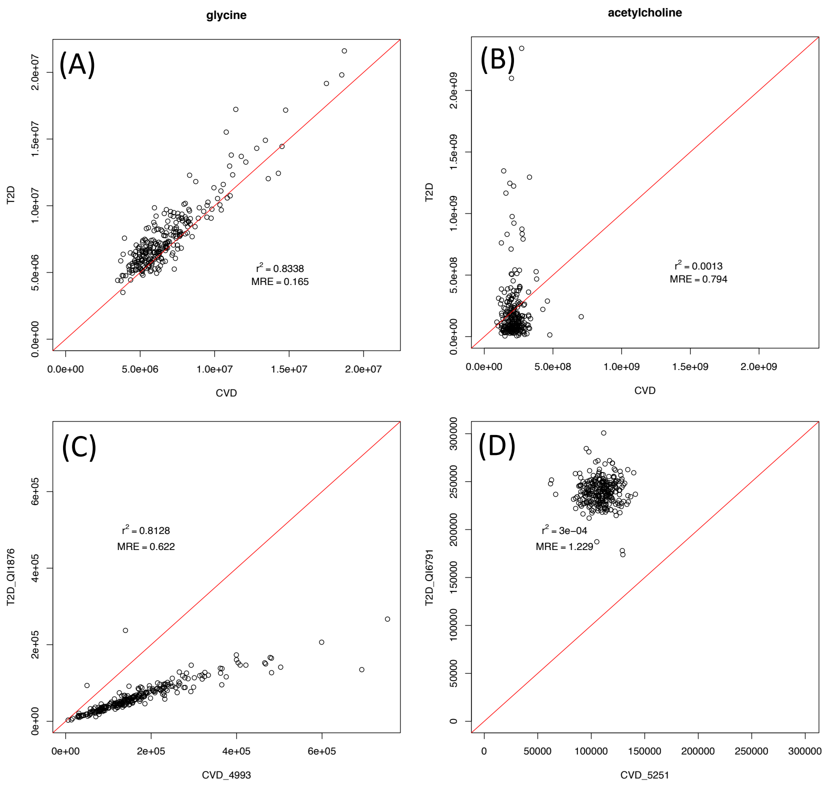


Figure 3.2. Distribution of selected known (A&B) and matched unknown (C&D) metabolites among the overlapping samples before normalization

X-axis: values in the CVD panel; y-axis: values in the T2D panel; axis labels for C&D indicate the assigned ID for the unknown metabolite in each study

Normalization results for the 83 known metabolites showed that the four methods performed almost equally well for most of the metabolites that had strong concordance to begin with. MRE was decreased and r^2 was either remained the same or slightly increased among the overlapping samples (Figure 3.3; top) While for metabolites that had a very skewed distribution in one panel but not the other, only quantile normalization worked to improve the correlation as well as reduce the MRE, scaling the extreme outliers in the non-reference panel back to the distribution of the reference panel. Median normalization had almost no effect, and robust regression was fitted poorly and failed to retain the relative magnitude of metabolite values (Figure 3.3; bottom). This is more evident in Figure 3.4, where the two quantile normalization methods were shown to reduce MREs the most, especially for metabolites that were the least consistent among the shared samples before normalization. Median normalization had only limited effect on improving concordance, and robust regression could sometimes enlarge the cross-study difference (Figure 3.4).

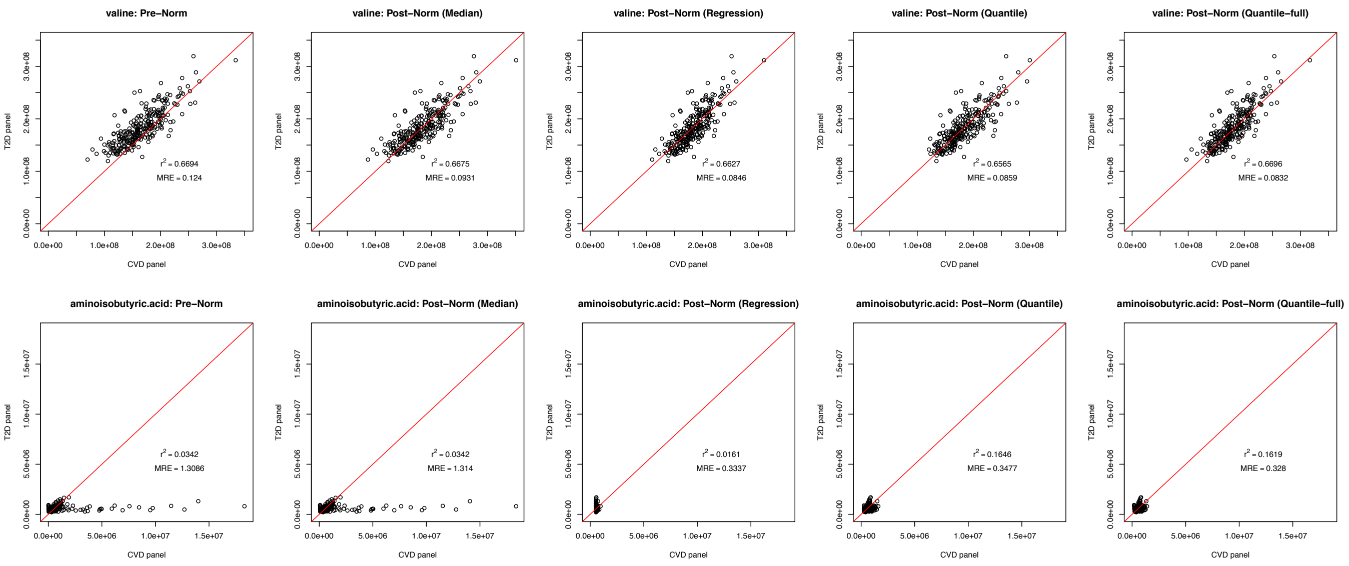


Figure 3.3. Distribution of selected known metabolites (top: valine; bottom: aminoisobutyric acid) among the overlapping samples after normalization comparing the four different methods

Left to right: the distributions before normalization, after "median normalization", after "robust regression", after "quantile normalization based on overlapping samples", and after "quantile normalization based on the full sample"; each point is a pair of the orignal observation in the reference panel vs. the average normazlied value in the non-reference panel

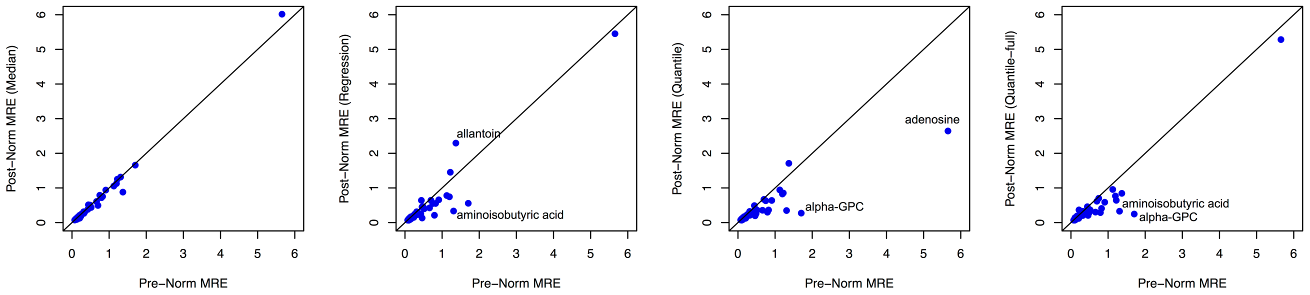


Figure 3.4. Change in MRE after normalization for the 83 known metabolites

X-axis: MRE before normalization; y-axis: *average* post-normalization MRE; left to right: median normalization, robust regression, quantile normalization based on overlapping samples, and quantile normalization based on the full sample

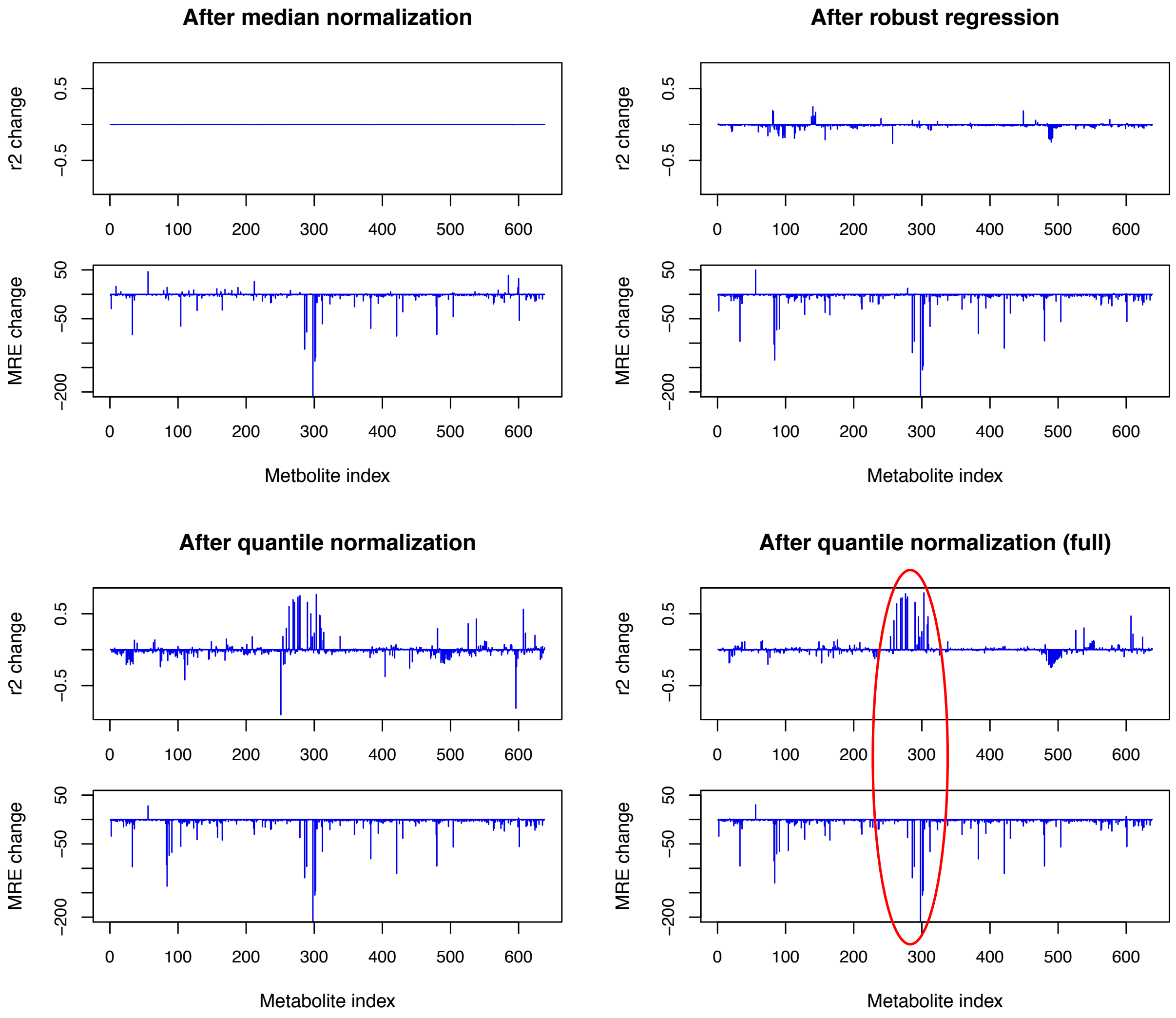


Figure 3.5. Change in MRE and r^2 after normalization for the 638 matched unknown metabolites

Y-axis: difference between post-normalization r^2 or MRE and the pre-normalization value

To examine the normalization results for all 638 matched unknown metabolites, we screened over the change in MRE and r^2 pre- and post-normalization for each method. The results showed that, while all methods worked to reduce MRE for most of the metabolites, quantile normalization based on all samples outperformed the other methods in increasing r^2 for a cluster of the metabolites (red circle; figure 3.5). Metabolites inside the cluster often had a few extreme outliers in one panel but not the other, which dragged down the correlation and concordance among the shared samples (Figure 3.6; top). Situations like this can be rescued by quantile normalization but not the other methods under comparison. For over half of the matched unknown metabolites that had a moderate to strong between-panel correlation but had systematically higher values in one of the panels, all methods except median normalization performed well in shifting and aligning the target distribution toward the other (Figure 3.6; bottom). For metabolites that were barely compatible across panels and were matched likely owing to noise (Figure 3.2D), none of the normalization methods could work to improve its correlation among the shared samples. Pre-ICC for these metabolites was typically close to -1; post-ICC was below or close to 0. The number of matched unknown metaoblties with an ICC > 0.4 increased from 134 before normalization to 512 after quantile normalization based on the full sample. These top, reliable matched metabolites could serve as candidates for experimental examinations to enhance the identification and understanding of unknown metabolites.

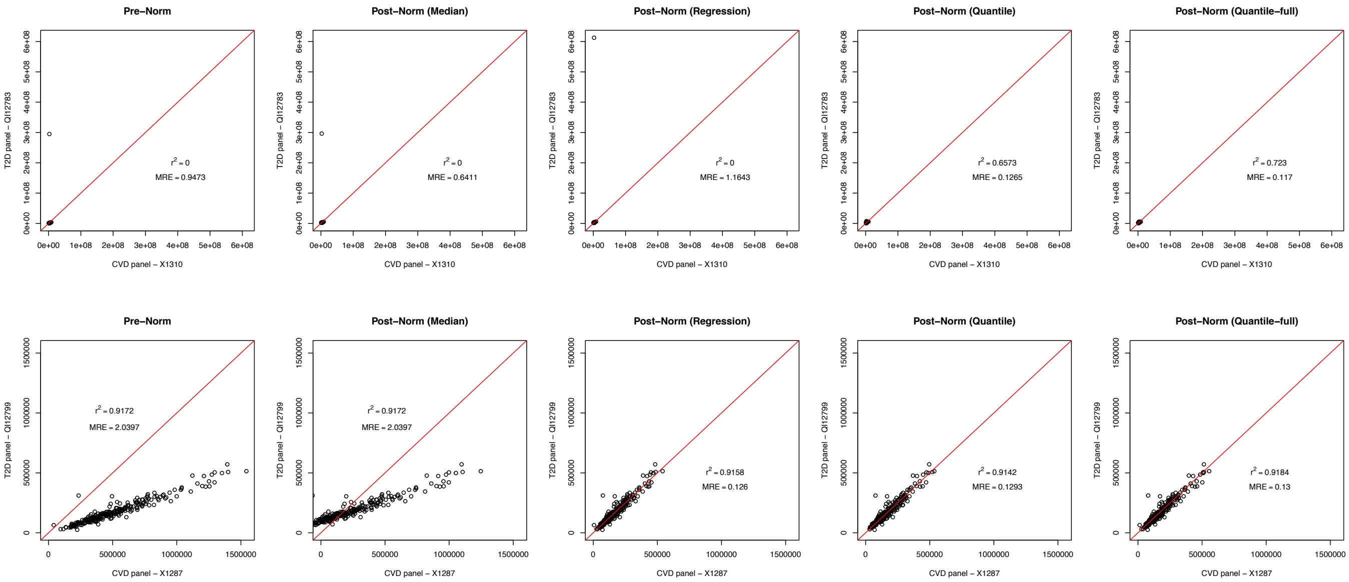


Figure 3.6. Distribution of selected matched unknown metabolites among the overlapping samples after normalization

Left to right: the distributions before normalization, after “median normalization”, after “robust regression”, after “quantile normalization based on overlapping samples”, and after “quantile normalization based on the full sample”; each point is a pair of the orignal observation in the reference panel vs. the average normazlied value in the non-reference panel; axis labels indicate the assigned ID for the unknown metabolite in each study

Discussion

We described here a strategy to calibrate metabolites across studies that utilizes the cross-study difference in metabolite distribution among overlapping samples as a function to align displaced distributions with one another. We implemented the idea via median normalization, robust regression, and quantile normalization, and examined their performance in normalizing values for both targeted and untargeted metabolites among 260 identical biological samples shared across two metabolomics datasets.

Results suggest quantile normalization as the preferred method over the most common median normalization or robust regression for cross-study normalization, especially when the range of metabolite values differ a lot between studies, or in the presence of extreme outliers in one study but not the other(s). In addition, using the empirical distribution from all samples for quantile normalizaiton provides a better resolution than using that based on the overlapping samples alone, as it covers the range of the full sample to which the normalization function would eventually apply. Another advantage of quantile normalization based on all samples is that it does not depend on having overlapping samples across studies, therefore can be readily applicable to datasets without shared samples.

Quantile normalization outperformed the other methods on the motivation datasets under study, but it may not always be the optimal approach given different datasets can have different data structure or metabolite distribution, particularly when generated from different platforms. Each normalization method has its own merits. Median normalization is easy and intuitive, and can effectively correct for subtle cross-study difference when concordance of metabolites among overlapping samples is strong in the first place. Robust regression preserves the paired information among overlapping samples in different studies, as metabolite values in the non-reference distribution are calibrated according to its one-to-one relationship with values in the reference distribution. Quantile normalization does not rely on and so does not necessarily retain the paired relationship after normalization. A better approach would be to incorporate the paired information of overlapping samples across studies into quantile normalization when calibrating the between-study batch effect.

One potential concern of quantile normalization is that it might break down the metabolite-phenotype association when the proportion of cases differs a lot between studies. For example, when one dataset contains 100% cases, and the other dataset contains 0% cases and 100% controls, we would expect metabolites that have a causal relationship with the outcome to have different distributions between the two datasets. Using quantile normalization to normalize values across the two datasets would force the distribution of the cases and that of the controls identical, hence losing power to detect associations of metabolites with the outcome of interest. This is not a unique problem and have been discussed in gene expression studies [17, 18]. While it is not very likely to have such an extreme case in real life, a possible solution to this can be to use a set of "housekeeping" metabolites [19, 20], just as housekeeping genes, whose level does not change with disease status, to construct the empirical distribution and the quantile function for mapping other non-housekeeping metabolites to correct for the cross-study deviation.

In conclusion, we proposed an idea for cross-study normalization that uses the information from overlapping samples and identified quantile normalization as the most effective method in adjusting the cross-study variation for both targeted and untargeted metabolites. This approach can benefit many of the existing pilot metabolomics studies with shared samples to increase their sample size and power for studying metabolite associations with a range of exposures and outcomes. The idea and approach illustrated in this work can be easily extended to more than two studies, and potentially to datasets generated from different platforms for a wider application.

Bibliography

1. Haznadar M, Maruvada P, Mette E, Milner J, Moore SC, Nicastro HL, Sampson JN, Su LJ, Verma M, Zanetti KA. Navigating the road ahead: addressing challenges for use of metabolomics in epidemiology studies. Metabolomics. 2014;10:176-178.
2. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol. 2012;13:263-269.
3. Clish CB. Metabolomics: an emerging but powerful tool for precision medicine. Cold Spring Harb Mol Case Stud. 2015;1:a000588.
4. Tzoulaki I, Ebbels TM, Valdes A, Elliott P, Ioannidis JP. Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. Am J Epidemiol. 2014;180:129-139.
5. Gika HG, Theodoridis GA, Wilson ID. Liquid chromatography and ultra-performance liquid chromatography-mass spectrometry fingerprinting of human urine: sample stability under different handling and storage conditions for metabonomics studies. J Chromatogr A. 2008;1189:314-322.
6. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, Lewis GD, Fox CS, Jacques PF, Fernandez C, O'Donnell CJ, Carr SA, Mootha VK, Florez JC, Souza A, Melander O, Clish CB, Gerszten RE. Metabolite profiles and the risk of developing diabetes. Nat Med. 2011;17:448-453.
7. Zelena E, Dunn WB, Broadhurst D, Francis-McIntyre S, Carroll KM, Begley P, O'Hagan S, Knowles JD, Halsall A, Consortium H, Wilson ID, Kell DB. Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. Anal Chem. 2009;81:1357-1364.
8. Want EJ, Wilson ID, Gika H, Theodoridis G, Plumb RS, Shockcor J, Holmes E, Nicholson JK. Global metabolic profiling procedures for urine using UPLC-MS. Nat Protoc. 2010;5:1005-1018.
9. Wang SY, Kuo CH, Tseng YJ. Batch Normalizer: a fast total abundance regression calibration method to simultaneously adjust batch and injection order effects in liquid chromatography/time-of-flight mass spectrometry-based metabolomics data and comparison with current calibration methods. Anal Chem. 2013;85:1037-1046.
10. Kirwan JA, Weber RJ, Broadhurst DI, Viant MR. Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control. Sci Data. 2014;1:140012.
11. Estruch R, Ros E, Salas-Salvado J, Covas MI, Corella D, Aros F, Gomez-Gracia E, Ruiz-Gutierrez V, Fiol M, Lapetra J, Lamuela-Raventos RM, Serra-Majem L, Pinto X, Basora J, Munoz MA, Sorli JV, Martinez JA, Martinez-Gonzalez MA. Primary prevention of cardiovascular disease with a Mediterranean diet. N Engl J Med. 2013;368:1279-1290.

12. Martinez-Gonzalez MA, Corella D, Salas-Salvado J, Ros E, Covas MI, Fiol M, Warnberg J, Aros F, Ruiz-Gutierrez V, Lamuela-Raventos RM, Lapetra J, Munoz MA, Martinez JA, Saez G, Serra-Majem L, Pinto X, Mitjavila MT, Tur JA, Portillo MP, Estruch R. Cohort profile: design and methods of the PREDIMED study. Int J Epidemiol. 2012;41:377-385.

13. Vazquez-Fresno R, Llorach R, Urpi-Sarda M, Lupianez-Barbero A, Estruch R, Corella D, Fito M, Aros F, Ruiz-Canela M, Salas-Salvado J, Andres-Lacueva C. Metabolomic pattern analysis after mediterranean diet intervention in a nondiabetic population: a 1- and 3-year follow-up in the PREDIMED study. J Proteome Res. 2015;14:531-540.

14. Wang DD, Toledo E, Hruby A, Rosner BA, Willett WC, Sun Q, Razquin C, Zheng Y, Ruiz-Canela M, Guasch-Ferre M, Corella D, Gomez-Gracia E, Fiol M, Estruch R, Ros E, Lapetra J, Fito M, Aros F, Serra-Majem L, Lee CH, Clish CB, Liang L, Salas-Salvado J, Martinez-Gonzalez MA, Hu FB. Plasma Ceramides, Mediterranean Diet, and Incident Cardiovascular Disease in the PREDIMED Trial. Circulation. 2017.

15. Zheng Y, Hu FB, Ruiz-Canela M, Clish CB, Dennis C, Salas-Salvado J, Hruby A, Liang L, Toledo E, Corella D, Ros E, Fito M, Gomez-Gracia E, Aros F, Fiol M, Lapetra J, Serra-Majem L, Estruch R, Martinez-Gonzalez MA. Metabolites of Glutamate Metabolism Are Associated With Incident Cardiovascular Events in the PREDIMED PREvencion con DIeta MEDiterranea (PREDIMED) Trial. J Am Heart Assoc. 2016;5.

16. Chaffin MD, Clish CB, Hu FB, Martínez-González MA, Bullo M, Corella D, Gómez-Gracia E, Fiol M, Estruch R, Lapetra J, Fitó M, Arós F, Serra-Majem L, Ros E, Liang L. MetProc: separating measurement artifacts from true metabolites in an untargeted metabolomics experiment. (submitted). 2016.

17. Qin S, Kim J, Arafat D, Gibson G. Effect of normalization on statistical and biological interpretation of gene expression profiles. Front Genet. 2012;3:160.

18. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733-739.

19. Ramirez T, Daneshian M, Kamp H, Bois FY, Clench MR, Coen M, Donley B, Fischer SM, Ekman DR, Fabian E, Guillou C, Heuer J, Hogberg HT, Jungnickel H, Keun HC, Krennrich G, Krupp E, Luch A, Noor F, Peter E, Riefke B, Seymour M, Skinner N, Smirnova L, Verheij E, Wagner S, Hartung T, van Ravenzwaay B, Leist M. Metabolomics in toxicology and preclinical research. ALTEX. 2013;30:209-225.

20. Lopez M, Lelliott CJ, Vidal-Puig A. Hypothalamic fatty acid metabolism: a housekeeping pathway that regulates food intake. Bioessays. 2007;29:248-261.