Published in final edited form as:

Wiley Interdiscip Rev Syst Biol Med. 2013 November ; 5(6): 677–686. doi:10.1002/wsbm.1242.

Systems Biology Approaches to Epidemiological Studies of Complex Diseases

Hongzhe Li

Professor of Biostatistics, Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

Abstract

Systems biology approaches to epidemiological studies of complex diseases include collection of genetic, genomic, epigenomic and metagenomic data in large-scale epidemiological studies of complex phenotypes. Designs and analyses of such studies raise many statistical challenges. This paper reviews some issues related to integrative analysis of such high dimensional and inter-related data sets and outline some possible solutions. I focus my review on integrative approaches for genome-wide genetic variants and gene expression data, methods for joint analysis of genetic and epigenetic variants and methods for analysis of microbiome data. Statistical methods such as mediation analysis, high dimensional instrumental variable regression, sparse signal recovery and compositional data regression provide potential frameworks for integrative analysis of these high dimensional genomic data.

INTRODUCTION

The field of genetic epidemiology, which focuses on the study of the role of genetic factors and environmental factors in determining health and disease in families and in populations, has shifted from family-based linkage studies to population-based studies of almost all common variants in the human genome. Although these studies have led to identification of many genetic variants that are associated with many complex diseases such as cardiovascular diseases, cancers and psychiatric disorders, these variants were identified in genome-wide association studies (GWAS) mainly through their marginal effects on disease risk. However, the mechanisms by which these genetic variants affect the phenotypes are still lacking for most of these complex diseases. This is further complicated by the fact that many such complex diseases are also due to behavioral and environmental factors. Simply collecting data of genetic variants and clinical phenotypes such as in traditional GWAS is not capable of fully addressing the epistasis and other interactions among the genetic variants and epigenomic variants that predominate many complex diseases. It is also becoming clear that a full understanding of gene-environment interactions requires that epigenetic mechanisms are taken into account,¹ including DNA methylation and histone modification that serve to regulate gene expression without altering the underlying DNA sequence. Adding to these complexities are the recent observations that human microbiota also play an important role in mediating the environmental factors and are associated with risk of many complex diseases.^{2, 3, 4, 5}

Systems biology approaches for epidemiological studies of complex diseases require collection of many intermediate phenotypes such as gene, protein and metabolite expression data and data related to epigenomics and microbiomes. Genetic and environmental factors influence clinical phenotypes by perturbing molecular networks in human cells and by perturbing gut microbiome composition. Systems-based approaches to epidemiological studies have the potential to interrogate these molecular phenotypes and identify patterns associated with disease. Epigenomics add another layer of complexity in gene regulation and

disease risk. More integrative, systems-based approaches for complex disease studies will be essential for large-scale genetic epidemiological studies. Studies that integrate multiple types of genomic data at the population level have the potential to reveal the underlying network of genes that drive disease initiation and progression.⁶ Powerful sequencing technologies have made collecting these diverse data on large sets of samples possible. However, novel statistical methods for integrating these diverse types of high dimensional genomic data are greatly needed.

In this paper, I review some methods for systems biology approaches to genetic/epigenetic epidemiological studies of complex phenotypes, focusing mainly on design and statistical analysis issues. I also outline several possible approaches that one can take for such integrative analyses, including mediation analysis,⁷ high dimensional instrumental variable regression,⁸ sparse shared signal recovery and high dimensional compositional data regression. These methods provide new statistical frameworks for analyzing complex high dimensional genomic/epigenomic data.

INTEGRATIVE ANALYSIS OF GENETIC VARIANTS, MOLECULAR PHENOTYPES AND COMPLEX DISEASES

One way to increase the power of identifying disease-associated variants is to incorporate other high-throughput functional datasets or molecular phenotype data such as genome transcriptome, proteome and microRNAome. The key issue and potentially limiting factor here is to collect these genome-wide functional data in relevant tissues that the disease expresses on. Such tissue specific data sets are now being generated and deposited into public databases (e.g., The Genotype-Tissue Expression (GTEx) project⁹). The ideal datasets for such an integrative analysis are SNP data and functional data all collected on the same set of samples (e.g., cases and controls). However, this is often not possible because of the cost or because the control samples simply do not have the appropriate tissues to study. The second type of datasets would be SNP data on all samples, but only a small subset of these samples will have genome-wide functional data. The third type of data is GWAS data and the functional data sets that are available on two sets of individuals. Different questions can be answered with different types of data sets with different statistical approaches.

Using eQTL Data to Increase the Power of Detecting Phenotype-Associated Variants - Filtering Approach

A simple approach to integrate high-throughput functional datasets (e.g., from studies of the transcriptome, proteome, or metabolome) with genome-wide genotype data is to select SNPs that meet certain functional criteria as determined by functional data sets such as gene expression and proteome expression data. The rationale for such an approach is the observation that trait-associated SNPs are more likely to be eQTLs.¹⁰ For example, Gamazon et al.¹¹ described a multi-step filtering approach, where in the first step, SNPs can be filtered by requiring that they be associated with genes whose expression levels are associated with the phenotypes. In the next step, the number of SNPs can be further reduced by requiring that they be associated with protein levels that are themselves associated with the disease. This process can also be applied to other molecular datasets. Similarly, other functional data such as those from the ENCODE project¹² can also be used to filter the relevant variants. The effect of such filtering steps is the identification of potentially more relevant SNPs and the reduction of the number of tests, and therefore can lead to an increase in power. If these filtering steps are done on independent data sets from the large GWAS data, one can simply perform multiple comparison adjustments on the set of SNPs that have passed the filtering steps. If the filtering samples overlap with the main GWAS samples, one has to adjust for the fact that the phenotype data are used in the filtering steps.¹¹

A Model-based Approach to Increase the Power of Detecting Disease-Associated Genetic Variants

If the SNPs and functional genomic data are collected on the same set of individual samples, a formal model-based approach can be used to identify the phenotype-associated SNPs taking into account the gene expression information. For the i^{th} subject, $i = 1, \dots, n$, let $Z_{ik} = 0, 1, 2$ be the number of minor alleles at the k^{th} SNP for $k = 1, \dots, K$, X_{ij} , $j = 1, \dots, p$ be the expression of the j^{th} transcript for $j = 1, \dots, p$, and W_{il} , $l = 1, \dots, q$ be additional non-genomic covariates, such as clinical or environmental measurements. Let $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$, $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iK})^T$, and $\mathbf{W}_i = (W_{i1}, \dots, W_{iq})^T$. Finally, let Y_i be the disease status with $Y_i = 1, 0$ denoting whether the subject has the disease or not.

We can integrate outcome, transcript expression levels, and genotype data according to the notion that a SNP impacts disease probabilities through regulation of gene expression. In particular, to assess the effect of the k^{th} SNP on disease risk, Zhao et al.¹³ proposes the following two-stage model,

$$\begin{aligned} \text{logit } P(Y_i=1|\mathbf{X}_i, \mathbf{Z}_{ik}, \mathbf{W}_i) &= \alpha_0^{\text{int}} + \mathbf{X}_i^T \boldsymbol{\alpha}_0 + \mathbf{W}_i^T \boldsymbol{\xi}_0, & (\text{outcome model}) \\ \mathbf{X}_i^T \boldsymbol{\alpha}_0 &= \beta_0^{\text{int}} + Z_{ik}\beta_0 + \mathbf{W}_i^T \boldsymbol{\nu}_0 + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), & (\text{transcript model}) \end{aligned} \quad (1)$$

where the random error ε_i is independent of $\mathbf{Z}_i$ and $\mathbf{W}_i$. The *outcome model* describes the effect of $\mathbf{X}_i$ on disease probabilities, while the *transcript model* describes the regulation of $\mathbf{X}_i$ by Z_{ik} . Here α_0^{int} and β_0^{int} are the intercepts, $\boldsymbol{\alpha}_0$ measures the effects of gene expression on disease risk, and $\boldsymbol{\nu}_0$ measures the effects of genetic variants on gene expressions. Furthermore, $\boldsymbol{\xi}_0$ and $\boldsymbol{\nu}_0$ are the coefficients associated with the observed covariates. The transcript model is similar to models used in the analysis of expression quantitative trait loci. In this model setting, we are concerned not with regulation of individual transcripts, but of one particular linear combination of them. The effect of this linear combination on disease probability, controlling for other non-genomic covariates, is modeled by the outcome model. Based on this model, one can test for SNP-disease association by testing $H_0: \beta_0 = 0$ by assuming the SNP affects disease risk through affecting the gene expression levels.

Using Genetic Variants to Find Phenotype-associated Genes - Matching Patterns of eQTL and GWAS

The third approach, which does not require the data collected on the same set of samples, is based on matching the association patterns. He et al.¹⁴ proposed such an approach for identifying the trait associated genes. They reasoned that for a disease-associated gene, any genetic variation that perturbs its expression is also likely to influence the disease risk. Thus, the expression quantitative trait loci (eQTL) of the gene, which constitute a unique genetic signature, should overlap significantly with the set of loci associated with the disease. They further developed a computational algorithm using the Bayes factor (named Sherlock) to search for gene-disease associations from GWAS, taking advantage of independent eQTL data.

Figure 1 illustrates this idea. Suppose that we convert the association p -values into Z -scores for both GWAS association and gene expression association for each of the p SNPs, and denote these Z -scores as $(X_1, \dots, X_p)$ for GWAS association, and $(Y_1, \dots, Y_p)$ for gene expression association. The goal is to find genes with patterns of associated SNPs that match the patterns of the SNPs that show association with the phenotype. However, we do not need to set the stringent criterion for genome-wide significance. The key is that if we observe a very small set of SNPs that are associated with both the gene expression and the phenotypes, then this is unlikely due to random chance. He et al.¹⁴ developed a Bayesian model and

calculate the Bayes factor to measure the gene-phenotype association. Alternatively, we can consider the following simultaneous signal detection problem. Assume that $X_i \sim N(\mu_i, 1)$ and $Y_i \sim N(\lambda_i, 1)$ where $\mu = (\mu_1, \dots, \mu_p)$ and $\lambda = (\lambda_1, \dots, \lambda_p)$ are the mean vectors and both are very sparse. We aim to test the following null hypothesis

$$H_0: \{i: \mu_i \neq 0 \text{ and } \lambda_i \neq 0\} = \emptyset$$

$$H_a: \text{There exists at least one } i \text{ such that } \mu_i \neq 0 \text{ and } \lambda_i \neq 0.$$

One possible test statistic is $S = \max_{i=1}^p \min(|X_i|, |Y_i|)$. Under this framework, one can investigate how signal strengths and frequencies affect the power of detecting such signals that are shared by gene expression phenotype and disease phenotype.

Using Genetic Variants to Find Phenotype-associated Genes - An Instrumental Variable Approach

Suppose we have a quantitative trait or clinical phenotype y , a p -vector of gene expression levels $\mathbf{X}$, and a q -vector of numerically coded genotypes $\mathbf{Z}$. In reality, there may be a sufficient set of unobserved confounding phenotypes $\mathbf{w}$ that act as proxies for the long-term effects of environmental exposures and/or the state of the microenvironment of the cells or tissues within which the biological processes occur. These phenotypes are likely to have strong influences on gene expression levels while contributing substantially to the clinical phenotype. Figure 2 illustrates the confounding between $\mathbf{X}$ and Y with an example of six variables. If an ordinary regression analysis is to be applied, the effects of X_1 and X_2 on y would be seriously confounded by w , resulting in an effect modification for X_1 and a spurious association for X_2 .

One way of controlling for the confounding due to $\mathbf{w}$ is through the use of the genotype $\mathbf{Z}$ as instruments. This is the underlying idea of Mendelian randomization in observational studies.⁸ In order for $\mathbf{Z}$ to be valid instruments, the following conditions must be satisfied:⁸

1. The genotype $\mathbf{Z}$ is (marginally) independent of the confounder $\mathbf{w}$;
2. The genotype $\mathbf{Z}$ is not (marginally) independent of the intermediate phenotype $\mathbf{X}$;
3. Conditionally on $\mathbf{X}$ and $\mathbf{w}$, the genotype $\mathbf{Z}$ and the response y are independent.

The above conditions are not easily testable from the observed data, but can be justified on the basis of plausible biological assumptions. Condition 1 is ensured by the usual assumption that the genotype is assigned at meiosis randomly, given the parents' genes, and independently of any possible confounder. Condition 2 requires that the genetic variants be reliably associated with the gene expression levels, which is often demonstrated by *cis*-eQTLs with strong regulatory signals. Condition 3 requires that the genetic variants have no direct effects on the clinical phenotype and can affect the latter only indirectly through the gene expression phenotypes. Owing to the large pool of gene expressions included in genetical genomics studies, the possibility of a strong indirect effect is greatly reduced and hence this condition is also mild and tends to be satisfied in practice.

Suppose we have n independent observations of $(y, \mathbf{x}, \mathbf{z})$. Denote by $\mathbf{y}$, $\mathbf{X}$, and $\mathbf{Z}$, respectively, the $n \times 1$ response vector, the $n \times p$ covariate matrix, and the $n \times q$ genotype matrix. Using the genotypes as instruments, we consider the following linear instrumental variables (IV) model for the joint modeling of the data $(\mathbf{y}, \mathbf{X}, \mathbf{Z})$:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\eta}, \\ \mathbf{X} &= \mathbf{Z}\boldsymbol{\Gamma}_0 + \mathbf{E}, \end{aligned} \quad (2)$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\Gamma}_0$ are a $p \times 1$ vector and a $q \times p$ matrix, respectively, of regression coefficients, and $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$ and $\mathbf{E} = (\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n)^T$ are an $n \times 1$ vector and an $n \times p$ matrix, respectively, of random errors such that the $(p+1)$ -vector $(\boldsymbol{\varepsilon}_i^T, \eta_i)$ is multivariate normal conditional on $\mathbf{Z}$ with mean zero and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{jk})$. We write $\sigma_{jj} = \sigma_j^2$. Without loss of generality, we assume that each variable is centered about zero so that no intercept terms appear in (2), and that each column of $\mathbf{Z}$ is standardized to have L_2 norm $\sqrt{n}$. We emphasize that $\boldsymbol{\varepsilon}_i$ and η_i may be correlated because of the arbitrary covariance structure. In contrast to the ordinary linear model regressing $\mathbf{y}$ on $\mathbf{X}$, model (2) does not require that the covariate $\mathbf{X}$ and the error $\boldsymbol{\eta}$ be independent, thus substantially relaxing the assumptions of ordinary regression models and being more appealing in data analysis. Wei et al.¹⁵ developed two-stage penalized estimation procedure to estimate the parameters and to simultaneously identify the possible instruments and genes that are associated with the phenotype $\mathbf{y}$.

eQTL Analysis in the Era of Next Generation Sequencing

Next generation sequencing technologies bring new opportunities and challenges to eQTL analysis and complex traits. First, whole genome sequencing can provide all rare and common variants of given individuals. Second, the RNA-seq technologies provide more detailed information about gene transcription, including whole gene expression and isoform-specific gene expressions and differential exon usages. One can then study the effects of both rare and common variants on gene regulations and can identify allele-specific gene expressions.^{16, 17}

INCORPORATING EPIGENOME INTO EPIDEMIOLOGICAL STUDIES

The role of epigenomics in gene regulation has been extensively studied in recent years.^{18, 19} Genes are packaged into chromatin and dynamic chromatin remodeling processes are required for the initial step in the gene transcription process, which is achieved by altering the accessibility of gene promoters and regulatory regions. Epigenetic factors, including DNA methylation, histone modifications, and the action of small non-coding RNAs such as microRNAs are responsible for this regulatory process. Chromatin structures hinder or allow the binding of transcription factors, which in turn determines the resulting gene expression patterns.²⁰ Because of the importance of epigenomics, an integrated epigenetic and genetic approach to common human diseases has also been articulated in recent reviews.^{1, 21, 18} Several large-scale epigenomic projects are underway and the data sets are expected to be available very soon. Examples include a 5-year ROADMAP Epigenomics Project of the US NIH, which focuses on 261 embryonic stem cell lines, fetal tissue and adult cells and tissues and 39 assays, including ChIP-seq for 30 histone modifications. Other nations and groups are doing similar things, some via the International Human Epigenome Project (IHEC). A recent large-scale twin study is proposed in the UK (TwinsUK)²² to discover methylated genes responsible for discordance of ten common traits and diseases, where epigenomic differences will be studied in 5000 adult UK twins aged 16–85, discordant and concordant for a wide variety of diseases and environments. Wong et al.²³ reported the first large-scale study to examine the role of genome-wide DNA methylation in autism spectrum disorder (ASD) and ASD-related traits. They observed that ASD-associated DNA methylation differences at numerous CpG sites, with some differentially methylated regions (DMRs) consistent across all discordant twin pairs.

Technologies and Data Available for Epigenomic Studies

The most commonly used Illumina Infinium 450k DNA Methylation Beadchip allows researchers to interrogate more than 485,000 methylation sites per sample at single-nucleotide resolution. It covers 99% of RefSeq genes, with an average of 17 CpG sites per gene region distributed across the promoter, 5'UTR, first exon, gene body, and 3'UTR. It covers 96% of CpG islands, with additional coverage in island shores and the regions flanking them. For each CpG site of interest, the array measures signal for methylated (M) and unmethylated (U). The consensus methylation level (β value B) at the i th CpG site is estimated as $B_i = \max(M_i, 0) / (\max(M, 0) + \max(U_i, 0) + e)$, where a small term $e = 100$ is added to avoid a very small number in the denominator. Similarly, log-ratios of methylated to unmethylated (M/U) signal can also be used. Much current research has focused on normalization of two different probe types (Inf I and Inf II), since the β values of the type I and type II probes distribute very differently. Teschendorff et al.²⁴ developed a Beta mixture quantile dilation (BMIQ) normalization algorithm to adjust the beta-values of type II design probes into a statistical distribution characteristic of type I probes. In addition, the presence of CpG sites and presence of SNPs in probe can affect the observed signal. Methods are in active development for going from differentially methylated sites in methylation variable positions (MVP) to differentially methylated regions (DMRs).

Next-generation sequencing-based technologies enable DNA methylation profiling at high resolution and low cost.²⁵ These methods can be broadly classified as methylation analysis of CpG enriched regions and whole genome methylation analysis. Examples of CpG enrichment methods include Methyl-Seq²⁶ and reduced representation bisulfite sequencing (RRBS).²⁷ Another most commonly used technique for profiling methylation is MeDiP-Seq. This technique is done by immunoprecipitation of methylated cytosines and subsequent sequencing.²⁸ In contrast to CpG enrichment methylation analysis, whole-genome bisulfite sequencing offers the ability to measure absolute levels of DNA methylation at single nucleotide resolution, but it is expensive because it requires sequencing of whole genomes. Lister et al.²⁹ presented such an approach. Specifically, for sample i , let n_{it} be the total number of reads from cytosine t and X_{it} be the number of methylated reads from cytosine t . The CpG-level summary is simply the proportion X_{it}/n_{it} . Each X_{it} follows a binomial distribution with success probability π_{it} , which represents the true proportion of cells for which the t th CpG is methylated in the sample i . The success probability can be estimated by $\hat{\pi}_{it} = X_{it}/n_{it}$. Assuming π_{it} is a smooth function of t along the genome, Hansen et al.³⁰ proposed to use local likelihood smoothing to estimate the methylation level in a genomic region for a single sample. After smoothing, they proposed to identify the differential methylated CpG and the differentially methylated regions based on two-sample t -statistics.

The N-terminal tails of histones are extensively modified in response to developmental and environmental signals. Histone modifications have important roles in transcriptional regulation, DNA repair, DNA replication, alternative splicing and chromosome condensation.³¹ The predominant method for mapping these post-translational modifications genome-wide involves a technique known as chromatin immuno-precipitation (ChIP). The powerful ChIP-seq methods are now routinely used for studying histone modifications. Many methods have been developed for single-sample ChIP-seq data analysis for identifying the regions with histone modifications, including the model-based analysis (MACS)³² which works well for identifying more localized regions and methods that are particularly designed for identifying broad domains such as the ChIP-seq data from histone modification.³³ Similarly, there are methods available for identifying the genomic regions with differential binding. Hidden Markov models have also been developed to combine information across multiple histone modification profiles in order to define the cellular states.³⁴

Different from genetic variants, which are largely fixed throughout the life course, epigenetic patterns not only vary from tissue to tissue but alter with advancing age and are sensitive to environmental exposures. Most of the genome-wide epigenetic studies, especially methylation studies, are based on unfractionated blood samples. The dynamic nature of the epigenetic data raises new statistical challenges in design and analysis of epigenome-wide association studies of common human diseases.³⁵ In particular, it is important to have a way of using the epigenomic data in analysis of GWAS.

Epigenomic Marks as Surrogates for Environmental Exposures

Epigenomic marks can also be used as surrogates for quantifying various environmental exposures.³⁶ The contribution made by environmental factors may be mediated through epigenetics. Because there can be multiple unknown environmental/behavior risk factors that predispose to disease, it can be difficult to collect all possible factors for association with the disease status. In addition, accurate measurements of environmental exposures such as physical activities and diet can be very difficult to obtain. Since the environment exposures up to the study time influence epigenetic states, these states are inherently quantifiable variables that can be used as surrogates for diverse unknown disease predisposing environmental factors.^{20, 37, 38, 36} As more and more data on how environments affect epigenomic states are available, it is possible to identify the specific genomic regions that have an enhanced epigenetic variability these nutritional (and other environmental) influences during development.³⁸ These epigenetics marks can be used to calibrate self-report assessments such as nutrient and physical activity and then to incorporate them into large-scale genetic association studies. The challenge is how to identify the epigenomic marks that can be used as surrogates for different environmental and behavioral factors and how to adjust these epigenomic marks in large-scale genetic epidemiological studies.

Suppose that we are interested in testing the association between a SNP coded as X_s and a phenotype Y , but we want to adjust for potentially p high dimensional epigenetic marks measured on disease relevant tissues, denoted by $Z_1, \dots, Z_p$. For continuous phenotype Y , we can assume that

$$Y = \beta_0 + \beta_s X_s + f(Z_1, \dots, Z_p) + \varepsilon,$$

where $f(Z_1, \dots, Z_p)$ is some pre-defined function to measure the possible effects of the epigenetic marks on Y . The null hypothesis of interest is $H_0 : \beta_s = 0$. We need a valid statistical test for this null hypothesis treating $(Z_1, \dots, Z_p)$ as potentially high dimensional confounding factors. When p is large, one should also select the relevant epigenetic marks in order to increase the power of such tests.

Joint Analysis of Genetic Variants and Epigenetic Variants

Another possible integration of epigenomic data into genetic association studies is to perform gene-based genetic-epigenetic tests. Specifically, for individual i , let Gv_{ig} be all the genetic variants in gene g , Me_{ij} be the vector of the measurements of methylation states at the CpG sites within and around gene g , and Hm_{ig} be the measurements of histone modifications in the promotor, gene-body and downstream of gene g (see Figure 3 for an illustration). In general, we can perform generalized linear regression analysis to link the phenotype to these gene-specific genetic and epigenetic data by modeling the mean function as

$$h(\mu_i) = \beta_0 + \beta_g Gv_{ig} + \beta_m Me_{ig} + \beta_h Hm_{ig} + \beta_{gm} Gv_{ig} Me_{ig} + \beta_{gh} Gv_{ig} Hm_{ig},$$

where h is the link function, β_g , β_m and β_h are the coefficients measuring the effects of genetic variants, methylation marks and histone modifications, and β_{gm} and β_{gh} are the coefficients associated with the interactions. These interaction terms can be used to model the epigenetic modification of disease penetrance. This model generalizes the simple multiplicative model considered in Slatkin.³⁹ When family data are available, one can study the loss or gain of particular epigenetic marks in generations.

INCORPORATING MICROBIOMES INTO EPIDEMIOLOGICAL STUDIES

Yet another layer of complication to epidemiological studies of complex phenotypes is the effect of microbiomes on human health and diseases. Such associations have been clearly demonstrated for complex diseases such as obesity,³ autoimmune disease⁴ and cardiovascular diseases.⁵ Many of these diseases may be related to changes in the microbial gene activities or composition of the gut microbiome, which has probably been profoundly affected by our lifestyle changes such as antibiotic use and diets over the last 50 or so years. Perturbation of microbiota is potentially dangerous and may be a root cause of many modern day chronic diseases.⁴⁰ In addition, host genetics and the environment can also shape the gut microbiota. These three factors may interact in the context of chronic disease.⁴¹

Next generation sequencing technologies have made it possible to survey all the microbial genomes of a given body site. There are usually two approaches to studying human microbiome. One approach is to sequence the marker genes such as 16S ribosome genes for studying bacterial composition at the genus level. Alternatively, one can sequence the DNAs directly from the environment samples to give a global characterization of all the microbiome, including both the microbial composition at the species/strain level and also the microbial gene compositions.^{42, 43} Several such studies have been published with the goal of establishing a set of reference data, such as catalogues of genes, microbial species and complete genome sequences of strains colonizing the various body sites, including the Human Microbiome Project (HMP)⁴⁴ in the US and the Metagenomics of the Human Intestinal Tract (MetaHIT) project in Europe.⁴⁵ Results from these studies have important implications for epidemiological studies of complex diseases.⁴⁶

Different from other types of genomic data, the final data set in microbiome studies is often the composition of the bacterial taxa, which is of high dimensional, but also very sparse. If we denote $X_i = (X_{i1}, \dots, X_{ip})$ as the relative abundance of the p bacterial taxa for sample i , we have

$$\sum_{k=1}^p X_{ik} = 1, X_{ik} \geq 0,$$

and $\sum_{k=1}^p I(X_{ik} \neq 0)$ is small while p is large since we do not expect to observe all the p taxa in a given sample. This raises many statistical challenges to analyze such sparse high dimensional compositional data. Important statistical questions include generalized linear regression analysis with X_i as covariates, principal component analysis of X_i and graphical models for such compositional data. New methods must simultaneously account for the compositional nature of the data and also the sparsity of the data. Methods are also needed

for differential abundance analysis with the goal to identify the bacterial species that are differentially abundant between cases and controls.

The log-contrast models were originally introduced by Aitchison and Bacon-Shone⁴⁷ for modeling experiments with mixtures, and has proved to be useful for a wide variety of regression problems with a composition playing the role of covariate. Suppose that we observe an n -vector y of responses and an $n \times p$ matrix $X = (X_{ij})$ of covariates, with each row of X lying in the $(p - 1)$ -dimensional positive simplex $S^{p-1} = \{(X_1, \dots, X_p): X_j > 0 \text{ for } j = 1, \dots, p \text{ and } \sum_{j=1}^p X_j = 1\}$. The linear log-contrast model can be specified as

$$Y = Z^p \beta_{\setminus p}^* + \varepsilon, \quad (3)$$

where $Z^p = (\log(X_{ij}/X_{ip}))$ is the $n \times (p - 1)$ log-ratio matrix with the p th component taken as the reference component,⁴⁸ $\beta_{\setminus p}^* = (\beta_1^*, \dots, \beta_{p-1}^*)^T$ is the corresponding $(p - 1)$ -vector of regression coefficients, and ε is an n -vector of independent noise distributed as $N(0, \sigma^2)$. By

introducing a new coefficient $\beta_p^* = -\sum_{j=1}^{p-1} \beta_j^*$, model (3) can be more conveniently expressed in the symmetric form

$$Y = Z \beta^* + \varepsilon, \quad \sum_{j=1}^p \beta_j^* = 0, \quad (4)$$

where $Z = (z_1, \dots, z_p) = (\log X_{ij})$ is the $n \times p$ design matrix and $\beta^* = (\beta_1^*, \dots, \beta_p^*)^T$ is the p -vector of regression coefficients. We are concerned with the high-dimensional sparse setting, where the dimensionality p is comparable to or much larger than the sample size n , while only a small portion of the regression coefficients are nonzero. Wei et al.⁴⁹ developed a regularization-based estimation procedure to select the phenotype-associated taxa and to estimate their effects in model (4).

DISCUSSION

Exploring interactions between the epigenome, inherited DNA sequence variation and micro-biome with the aim of undertaking an integrated genetic-epigenetic-metagenomic approach to diseases enables us to consider many of the factors that contribute to the susceptibility of complex disease. The idea of systems epidemiology has also been articulated in a few other recent reviews^{50, 51} and the power of such integrative approaches has been clearly demonstrated in recent studies. Ghazalpour et al.⁵² integrated genetic and gene expression network to identify the genetic targets that influence gene modules (pathways) that are related to mouse weight. Rhinn et al.⁵³ took an integrative genomics approach based on analysis of transcriptional networks in human brain to identify apolipoprotein E (APOE) $\epsilon 4$ effectors in late-onset Alzheimer's disease. Aran and Hellman⁵⁴ showed that the distal enhancer sites contain both sequence and methylation polymorphisms, but the association between sequence variants and gene expression levels of cancer genes is weak in estrogen receptor (ER)-positive breast tumors. They observed that methylation level, which integrates genetic and environmental clues, is better correlated with gene expression.

Modern sequencing technologies make it possible to collect these data on a very large set of samples. Integrative systems biology approaches to analysis of these data represent a major challenge and also opportunities for biostatisticians to develop novel analytical methods to simultaneously analyze multiple very high dimensional data. We need new statistical

frameworks that can simultaneously consider many different dimensions of data. With full genome sequence and epigenomic maps of the DNA methylation and modified histone landscapes, methods are needed to identify exactly which genes are “turned on” and in which tissues and to identify how the combinations of these factors contribute to disease risks.⁵⁵ These data sets also offer the detailed views of genetic and epigenetic regulatory networks and how perturbation of such networks leading to disease. Statistically, we need new formulations to describe network perturbation and the causes for such changes. Integration of this massive amount of data promises to revolutionize our understanding of gene-gene, gene-environment and microbiome-host interactions and offer new ways to diagnose, prevent and treat complex diseases.

Acknowledgments

I thank Drs. Wei Lin and Sihai Dave Zhao for helpful discussions. The research was supported by NIH grants CA127334 and GM097505.

References

1. Bjornsson HT, Fallin MD, Feinberg AP. An integrated epigenetic and genetic approach to common human disease. *Trends in Genetics*. 2004; 20:350–358. [PubMed: 15262407]
2. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nature Review Genetics*. 2012; 13:260–270.
3. Ley RE, Turnbaugh P, Klein S, Gordon JL. Human gut microbial ecology linked to obesity. *Nature*. 2006; 444:1022–1023. [PubMed: 17183309]
4. Markle JG, Frank DN, Mortin-Toth S, Robertson CE, Feazel LM, Rolle-Kampczyk U, von Bergen M, McCoy KD, Macpherson AJ, Danska JS. Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science*. 2013; 339:1084–1088. [PubMed: 23328391]
5. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, Dugar B, Feldstein AE, Britt EB, Fu X, Chung YM, Wu Y, Schauer P, Smith JD, Allayee H, Tang WH, DiDonato JA, Lusis AJ, Hazen LS. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*. 2011; 472:57–63. [PubMed: 21475195]
6. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE. Variations in DNA elucidate molecular networks that cause disease. *Nature*. 2008; 452:429–435. [PubMed: 18344982]
7. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*. 1986; 51:1173–1182. [PubMed: 3806354]
8. Sheehan, Nuala A.; Didelez, Vanessa; Burton, Paul R.; Tobin, Martin D. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Medicine*. 2008; 5(8):1205–1210.
9. The GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nature Genetics*. 2013; 45:580–585. [PubMed: 23715323]
10. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genetics*. 2010; 6:e1000888. [PubMed: 20369019]
11. Gamazon E, Huang R, Dolan E, Cox N, Im H. Integrative genomics: Quantifying significance of phenotype-genotype relationships from multiple sources of high-throughput data. *Frontiers in Genetics*. 2012 doi:10.3389:fgene.2012.00202.
12. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. [PubMed: 22955616]
13. Zhao, SD.; Cai, TT.; Li, H. Technical report. University of Pennsylvania; 2013. More powerful genetic association testing via a new statistical framework for integrative genomics.

14. He X, Fuller CK, Song Yi, Meng Q, Zhang B, Yang X, Li H. Sherlock: Detecting gene-disease associations by matching patterns of expression QTL and GWAS. *American Journal of Human Genetics*. 2013; 92:667–680. [PubMed: 23643380]
15. Lin, W.; Feng, R.; Li, H. High-dimensional instrumental variables regression for associating complex traits with genetical genomics data. 2013. Available at arxiv.org/pdf/1304.7829
16. Sun W. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics*. 2012; 68:1–11. [PubMed: 21838806]
17. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*. 2010; 464:768–772. [PubMed: 20220758]
18. Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. *Nature Reviews Genetics*. 2007; 8:253–262.
19. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics*. 2007; 8:286–298.
20. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*. 2003; 33:245–254. [PubMed: 12610534]
21. Laura B. Epigenomics: The new tool in studying complex diseases. *Nature Education*. 2008; 1:1.
22. Bell JT, Spector TD. A twin approach to unraveling epigenetics. *Trends in Genetics*. 2011; 27:116–125. [PubMed: 21257220]
23. Wong CCY, Meaburn EL, Ronald A, Price TS, Jeffries AR, Schalkwyk LC, Plomin R, Mill J. Methylomic analysis of monozygotic twins discordant for autism spectrum disorder and related behavioural traits. *Molecular Psychiatry*. 2013 advance online publication.
24. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics*. 29:189–96. 201. [PubMed: 23175756]
25. Lister R, Ecker JR. Finding the fifth base: Genome-wide sequencing of cytosine methylation. *Genome Research*. 2009; 19:959–966. [PubMed: 19273618]
26. Brunner AL, Johnson DS, Kim SW, Valouev A, Reddy TE, et al. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Research*. 2009; 19:1044–1056. [PubMed: 19273619]
27. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, et al. Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic Acids Research*. 2005; 33:5868–5877. [PubMed: 16224102]
28. Taiwo O, Beck S, Butcher LM, Wilson GA, Morris T, Seisenberger S, Reik W, Pearce D. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nature Protocols*. 2012; 7:617–636.
29. Lister R, OMalley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*. 2008; 133:523–536. [PubMed: 18423832]
30. Hansen KD, Langmead B, Irizarry RA. Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*. 2012; 13:R83. [PubMed: 23034175]
31. Portela A, Esteller M. Epigenetic modifications and human disease. *Nature Biotechnology*. 2010; 28:1057–1068.
32. Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*. 2008; 9:R137. [PubMed: 18798982]
33. Zang Z, Schones DE, Zeng Z, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*. 2009; 25:1952–1958. [PubMed: 19505939]
34. Ernst J, Kellis M. Chromhmm: automating chromatin-state discovery and characterization. *Nature Methods*. 2012; 9:215–216. [PubMed: 22373907]
35. Rakyen VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nature Review Genetics*. 2011; 12:529–541.

36. Maunakea AK, Chepelev I, Zhao K. Epigenome mapping in normal and disease states. *Circulation Research*. 2010; 107:327–339. [PubMed: 20689072]
37. Feil R. Environmental and nutritional effects on the epigenetic regulation of genes. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*. 2006; 600:46–57.
38. Waterland RA, Jirtle RL. Early nutrition, epigenetic changes at transposons and imprinted genes, and enhanced susceptibility to adult chronic diseases. *Nutrition*. 2004; 20:63–68. [PubMed: 14698016]
39. Slatkin M. Epigenetic inheritance and the missing heritability problem. *Genetics*. 2009; 182:845–850. [PubMed: 19416939]
40. Nicholson JK, Holmes E, Wilson ID. Gut microorganisms, mammalian metabolism and personalized health care. *Nature Review Microbiology*. 2005; 3:431–438.
41. Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Review Microbiology*. 2011; 9:279–290.
42. Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Research*. 2009; 19:1141–1152. [PubMed: 19383763]
43. Morgan XC, Huttenhower C. Chapter 12: Human microbiome analysis. *PLoS Computational Biology*. 2012; 8
44. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012; 486:207–214. [PubMed: 22699609]
45. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010; 464:59–65. [PubMed: 20203603]
46. Foxman B, Rosenthal M. Implications of the human microbiome project for epidemiology. *American Journal Epidemiology*. 2013; 177:197–201.
47. Aitchison J, Bacon-Shone J. Log contrast models for experiments with mixtures. *Biometrika*. 1984; 71:323–330.
48. Aitchison J. The statistical analysis of compositional data (with discussion). *Journal of Royal Statistical Society B*. 1982; 44:139–177.
49. Lin, W.; Shi, P.; Feng, R.; Li, H. Technical report. University of Pennsylvania; 2013. Variable selection in regression with compositional covariates.
50. Haring R, Wallaschofski H. Diving through the -omics: the case for deep phenotyping and systems epidemiology. *OMICS*. 2012; 16:231–244. [PubMed: 22320900]
51. Hu FB. Metabolic profiling of diabetes: from black-box epidemiology to systems epidemiology. *Clinical Chemistry*. 2011; 57:1224–1236. [PubMed: 21690202]
52. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genetics*. 2006; 2:e130. [PubMed: 16934000]
53. Rhinn H, Fujita R, Qiang L, Cheng R, Lee JH, Abeliovich A. Integrative genomics identifies APOE e4 effectors in Alzheimer's disease. *Nature*. 2013; 1038/nature1241
54. Aran D, Hellman A. Dna methylation of transcriptional enhancers and cancer predisposition. *Cell*. 2013; 154:11–13. [PubMed: 23827668]
55. van Steensel B. Mapping of genetic and epigenetic regulatory networks using microarrays. *Nature Genetics*. 2005; 37:S18–S24. [PubMed: 15920525]

Figure 1.

Matching patterns of GWAS signals and eQTL signals. Left plot: Z-scores for SNP-phenotype association; right plot: Z-score for eQTL association for a given gene transcript. Peaks labeled by * are those SNPs that are shared between phenotype and gene expression. The peak matching provides evidence for gene-phenotype association.

Figure 2.

A causal diagram showing the relationships between two genotypes z_1 and z_2 , two gene expression levels x_1 and x_2 , a clinical phenotype y , and a confounding phenotype w .

Figure 3.

Simultaneous consideration of all genetic variants (Gv_g), methylation states (Me_g) and histone modifications (Hm_g) for gene g in association analysis of complex phenotypes.