# Statistical Methods in Integrative Genomics

Sylvia Richardson,[1] George C. Tseng,[2] and Wei Sun[3,4]

[1]MRC Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge, Cambridge CB2 0SR, United Kingdom; email: sylvia.richardson@mrc-bsu.cam.ac.uk

[2]Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15261; email: ctseng@pitt.edu

[3]Department of Biostatistics, Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599; email: weisun@email.unc.edu

[4]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 27516

Annu. Rev. Stat. Appl. 2016. 3:181–209

First published online as a Review in Advance on April 18, 2016

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-041715-033506



## Keywords

genomics, integrative genomics, horizontal data integration, vertical data integration

## Abstract

Statistical methods in integrative genomics aim to answer important biology questions by jointly analyzing multiple types of genomic data (vertical integration) or aggregating the same type of data across multiple studies (horizontal integration). In this article, we introduce different types of genomic data and data resources, and then we review statistical methods of integrative genomics with emphasis on the motivation and rationale of these methods. We conclude with some summary points and future research directions.

## 1. INTRODUCTION

It is an exciting time to work on statistical methods for genomic problems. The rapid development of high-throughput techniques allows researchers to collect large amounts of genomic data that can answer biological questions and enable the development of more effective therapeutic strategies for human diseases. Because multiple types of genomic data are often available within and across studies, the integrated analysis of genomic data has become popular. One may integrate the same type of genomic data across multiple studies (horizontal integration) or integrate different types of genomic data in the same set of samples (vertical integration). We review both horizontal and vertical integration studies, putting more emphasis on the latter. Before discussing statistical methods, we give a brief review of different types of genomic data as well as resources on where to obtain genomic data and related annotations. Many of our discussions and analytical rationales are cancer-focused, although most of the discipline applies to diseases in general.

**Array comparative genomic hybridization (CGH):** CGH on DNA microarray to compare the copy number of two DNA samples

**Read depth:** the (average) number of times a nucleotide is covered by sequencing process

**Somatic DNA mutations:** DNA mutations that are not inherited from a parent or passed to offspring

**Tumor purity:** the percentage of tumor cells within a tumor sample

**Ploidy:** the number of sets of chromosomes within a cell. A normal human cell is diploid, with 2 sets of chromosomes

### 1.1. Different Types of Genomic Data

It is crucial to understand the characteristics of each type of genomic data used for the purposes of integrated genomics. Therefore we first review different types of genomic data.

**1.1.1. DNA.** A common type of genomic analysis is to study DNA features from germline (normal) tissue and tumor tissue separately, as they often have very different characteristics. DNA variants from germline tissue include single nucleotide polymorphisms (SNPs), indels (short insertions or deletions), copy number variations (CNVs), and other structural changes such as translocations. SNP arrays can provide high-confidence genotype estimates because the underlying genotypes belong to one of three classes: AA, AB, and BB, where A and B indicate the two alleles of a SNP. Most SNP arrays are designed to target common variants (DNA variants that occur in more than 1% of individuals in a population). Both array comparative genomic hybridization (CGH) and SNP arrays can measure CNVs. Whereas array CGH can only measure total copy number, SNP arrays can measure the allele-specific copy number, which is the copy number in each of two homologous chromosomes (Wang et al. 2007, Sun et al. 2009). Recently, high-throughput sequencing, including whole-genome sequencing or exome sequencing, has been used to study DNA variants. With sufficient read depth, sequencing data can provide more accurate estimates of SNP genotypes and copy number calls. In addition, sequencing data can detect rare mutations (mutations with low population frequencies) that are usually not captured by arrays (Mills et al. 2011, Nielsen et al. 2011).

In cancer studies, we are often interested in somatic DNA mutations that occur in tumor tissues but are not found in the germline. Somatic point mutations, including single nucleotide changes and indels, are often rare. It is likely that two cancer patients share few or no somatic point mutations across whole exonic regions. In this sense, cancer may be better considered as a collection of rare diseases rather than one disease. Owing to such rareness, somatic point mutations are usually detected by sequencing. A somatic copy number aberration (SCNA) often occupies a relatively long genomic region (up to one-third of a chromosome may be deleted or amplified) and can be relatively common. SCNAs can be studied by array CGH, SNP array, or high-throughput sequencing. Studying somatic DNA mutations (either point mutations or SCNAs) is challenging because tumor samples are often composed of a mixture of tumor and normal cells (e.g., the normal cells from connective tissues or blood vessels), and tumor cells may have more or fewer than two copies of DNA on average. These two issues are known as tumor purity and ploidy issues. Unknown purity and ploidy affect each other and should be estimated together (Van Loo

et al. 2010, Carter et al. 2012). In addition, recent sequencing studies have revealed that tumor cell populations may be composed of several tumor subclones. Some somatic mutations may only occur in one or some of the subclones and thus have low allele frequencies. It is challenging to distinguish such mutations from sequencing errors (Ding et al. 2014).

**Tumor subclone:** All tumor cells within a subclone are descended from the same cell and have the same set of somatic mutations

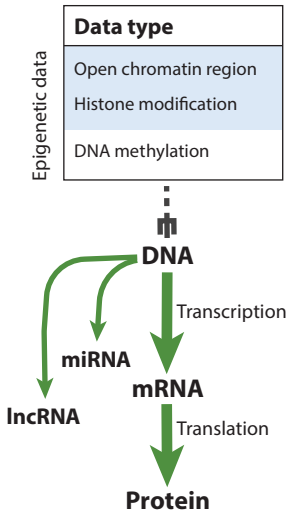
**1.1.2. Epigenetic marks.** Normal cells within the human body share almost identical DNA, with sporadic somatic mutations contributing a small amount of variation between cells. Despite this similarity, different types of cells are observed to have dramatically different sizes, shapes, and/or functions. Such cell-type-specific traits are often maintained by epigenetic marks, which are modifications on DNA molecules or proteins that can be passed to daughter cells during mitosis. The term epi means "over, outside of, around" in Greek. Although epigenetic marks do not change the DNA sequence itself, they may also be inheritable, and their role in the etiology of human diseases is increasingly recognized (Jiang et al. 2004). We introduce three types of epigenetic marks: open chromatin regions, histone modifications, and DNA methylation.

**Histone modifications:** methylation, acetylation, and phosphorylation of certain amino acids of histone proteins

**DNase I hypersensitive sites (DHSs):** genomic loci sensitive to cleavage by the DNase I enzyme where the DNA sequence is loosened and exposed instead of taking a condensed form

Within a cell nucleus, DNA is packed around multiple proteins called histones. This complex of DNA and proteins is referred to as chromatin. Chromatin usually takes a condensed form so that the packed DNA sequence is not accessible by other proteins such as regulatory transcription factors. Open chromatin regions, where previously packed DNA is loosened and exposed, often harbor active regulatory elements bound to DNA. Open chromatin regions can be detected by DNase I hypersensitive sites (DHSs) sequencing, where DNA sequences on DHSs are captured and then located by high-throughput sequencing techniques (**Figure 1**) (Song & Crawford 2010).

Histone modifications include different types of chemical modifications (e.g., methylation, acetylation, or phosphorylation) on different amino acids of histone proteins. Chromatin




| Data type | Platform | Data |
|---|---|---|
| Open chromatin region<br>Histone modification | DNase-seq<br>ChIP-chip/ChIP-seq | Amount of DNA sequence after enrichment by DNase or ChIP |
| DNA methylation | DNA methylation array<br>Bisulfite sequencing | Percent of methylated DNA sequence |
| | SNP array, exome-seq, whole-genome seq (WGS) | SNP genotype, indel calls |
| | Array CGH<br>Exome-seq, WGS | Intensity on probes<br>Sequence fragment counts |
| | Gene expression array | Array intensity on probes |
| | RNA-seq | Fragment counts reflecting RNA expression |
| | Protein array<br>Mass spectrometry | Measurement of the abundance of a protein or its phosphorylation forms |

**Example**

| Platform | ID | Sample 1 | Sample 2 |
|---|---|---|---|
| ChIP-seq | Locus 1<br>Locus 2 | 100<br>120 | 5<br>10 |
| Methylation array | CpG 1<br>CpG 2 | 0.95<br>0.95 | 0.6<br>0.72 |
| SNP array | SNP 1<br>SNP 2 | AG<br>CC | GG<br>CC |
| Exome-seq | Locus 1<br>Locus 2 | 50<br>75 | 20<br>32 |
| Expression array | Gene 1<br>Gene 2 | 3.2<br>6.3 | 3.4<br>5.4 |
| RNA-seq | Gene 1<br>Gene 2 | 521<br>856 | 232<br>387 |
| Protein array | Protein 1<br>Protein 2 | 0.34<br>−0.93 | 0.08<br>−1.1 |

**Figure 1**

Different types of genomic data and the platforms to measure these genomic data. The diagram on the left side illustrates the central dogma (*thick green arrows*), and that DNA may also encode other types of RNA (*thin green arrows*) such as microRNA (miRNA) or long noncoding RNA (lncRNA), and that DNA may be marked by different types of epigenetic marks. Other terms: CGH, comparative genomic hybridization; ChIP-chip, chromatin immunoprecipitation followed by microarray; ChIP-seq, chromatin immunoprecipitation followed by sequencing; DNase-seq, DNase I hypersensitive sites sequencing; exome-seq, exome sequencing; indel, short insertion or deletion; RNA-seq, RNA sequencing; SNP, single nucleotide polymorphism.

immunoprecipitation (ChIP) followed by microarray (ChIP-chip) or sequencing (ChIP-seq) is a popular choice to capture DNA sequences associated with modified histones (**Figure 1**). The ChIP step enriches for such DNA sequences, and the following microarray or sequencing step determines their likely genomic location. Each type of histone modification may occur on short or long genomic regions and is associated with certain biological features, such as active promoters or genes with suppressed expression (ENCODE Consortium 2012, Rashid et al. 2014).

DNA methylation usually refers to the addition of a methyl group to cytosine residues within CpG dinucleotides (nucleotide C followed by nucleotide G in a single strand of DNA). In the human genome, there are approximately 28 million CpG sites, which are not uniformly distributed. Clusters of CpG sites (called CpG islands) tend to occur on gene promoters (Stirzaker et al. 2014). DNA methylation on promoter regions usually represses gene expression; in contrast, DNA methylation in genic or exonic regions is often positively associated with gene expression. Popular techniques to measure DNA methylation including array-based methods [e.g., the Infinium HumanMethylation450 BeadChip array (HM450)], whole-genome bisulfite sequencing (WGBS), and reduced representation bisulfite sequencing (RRBS) (**Figure 1**). The HM450 provides two measurements for a CpG locus, methylation (M) and unmethylation (U) signals. A commonly used measurement of methylation is referred to as beta-value, which is equal to $M/(M + U)$ (see **Figure 1** for examples). Using WGBS, one can count the number of sequence reads with methylated or unmethylated CpGs, where methylated CpGs are marked by bisulfite transformation. Although RRBS covers less than 5% of CpGs genome-wide (~1 million of the 28 million CpG sites), its coverage is enriched for CpGs at promotor regions (~0.5 million of 2 million CpG sites on promoters) (Stirzaker et al. 2014).

**ChIP-chip/ChIP-seq:** Chromatin immunoprecipitation (ChIP) followed by microarray (chip) or sequencing

**DNA methylation:** an epigenetic mark of DNA sequence by adding a methyl group ($CH_3$) to DNA nucleotides

**Bisulfite sequencing:** sequencing DNA after bisulfite treatment that converts unmethylated cytosine to uracil while leaving methylated cytosine unaffected

**Microarray probes:** fragments of DNA sequence located in a spot of a microarray

**Phosphorylation:** the addition of a phosphate ($PO_4^{3-}$) group to a protein or other organic molecule

**1.1.3. RNA.** Three types of RNA molecules are commonly encountered in genomic data: messenger RNA (mRNA) that encode proteins, and two types of noncoding RNA with regulatory roles, microRNA (miRNA) and long noncoding RNA. The field has gradually recognized miRNA as a type of epigenetic machinery (Malumbres 2013). Expression of any type of RNA has traditionally been studied by different types of microarrays, where the expression of one gene/RNA may be measured by one or more microarray probes. In recent years, RNA sequencing (RNA-seq) has been replacing microarrays to become the major platform of transcriptomic studies. Compared with microarrays, RNA-seq provides more accurate estimates of gene expression, allows de novo discovery of transcripts, and delivers new information such as allele-specific expression and RNA isoform-specific expression. Recent studies have systematically evaluated different RNA-seq protocols and paved the way for future large-scale RNA-seq studies (Kratz & Carninci 2014). Using RNA-seq data, the expression of one gene could be quantified by the number of RNA-seq fragments mapped to this gene after correcting for read depth and gene length.

**1.1.4. Proteins.** Proteins perform many fundamental functions within living organisms, and their abundance and activity are important to understand. Protein expression, however, is often less studied, mostly because the amino acids do not form double-helix structures as in nucleotides, so the amplification and hybridization techniques used in microarray and sequencing for DNA and RNA cannot conveniently be applied to proteins. The activity of a protein may depend on a specific set of posttranslational modifications (PTMs). There are more than 200 types of PTMs that may occur in multiple positions in a protein, so the potential combinations of PTMs lead to an enormous number of protein states that cannot be handled by current technology. A particular form of PTM, phosphorylation, has been better studied because phosphorylated proteins (phosphoproteins) play important roles in signaling pathways, and assays are available to measure the large-scale abundance of phosphoproteins (Terfve et al. 2012). There are currently

**Table 1 Genomic databases and data access projects**

| Resource | URL | Full name (if applicable) and description |
|---|---|---|
| dbGAP | **http://www.ncbi.nlm.nih.gov/gap** | Database of Genotypes and Phenotypes; archive of human genotype and phenotype data |
| ArrayExpress | **http://www.ebi.ac.uk/arrayexpress** | Archive of functional genomics data including gene expression and epigenetic data |
| GEO | **http://www.ncbi.nlm.nih.gov/geo** | Gene Expression Omnibus; data repository for gene expression and epigenetic marks |
| SRA | **http://www.ncbi.nlm.nih.gov/sra** | Sequence Read Archive; database of sequencing data |
| TCGA | **http://tcga-data.nci.nih.gov/tcga/** | The Cancer Genome Atlas; portal to access multiple types of open-access genomic data from cancer samples |
| CGHub | **http://cghub.ucsc.edu** | Cancer Genomics Hub; repository for cancer genomic data that requires controlled access |
| ICGC | **http://dcc.icgc.org** | International Cancer Genome Consortium; portal to access multiple types of genomic data |
| Roadmap | **http://www.roadmapepigenomics.org** | Epigenomics project that provides human epigenomic data |
| GTEx | **http://www.gtexportal.org/home** | Genotype-Tissue Expression project; portal to access RNA sequence data from different tissues |
| ENCODE | **http://www.encodeproject.org** | Encyclopedia of DNA Elements; project to collect epigenetic and gene expression data from a diverse set of samples |



two classes of techniques for the study of proteomics: antibody-based arrays and mass spectrometry (MS). The Cancer Genome Atlas (TCGA) project has measured expression of more than 100 proteins or phosphoproteins across thousands of cancer patients using an antibody-based reverse phase protein array (RPPA). A traditional gene expression array measures genome-wide expression of one sample on an array; in contrast, each spot in an RPPA corresponds to a sample, and an RPPA measures the expression of one protein or phosphoprotein across all the samples spotted on this array. Therefore the output of an RPPA is comparable for one protein across all the samples, but in general is not comparable for multiple proteins within one sample. Proteomics is a fast-growing field. Several large-scale proteomics projects are ongoing, such as the Clinical Proteomic Tumor Analysis Consortium (Ellis et al. 2013).

**Mass spectrometry (MS):** a technique that measures the amount of analytes (e.g., protein peptides) by their mass-to-charge ratios

**Reverse phase protein array:** a protein array designed to measure the expression of one protein across multiple samples

## 1.2. Genomic Data Resources

A huge amount of publicly available genomic data is available in different databases (**Table 1**). The data from genome-wide association studies (GWAS), including DNA genotype and phenotype data, are often deposited in the dbGAP (Database of Genotypes and Phenotypes) (**Table 1**), which is hosted by National Center for Biotechnology Information (NCBI). Because genotyping information can theoretically trace to patient identity, one needs to complete a secure access application through dbGAP to protect privacy of patients. Gene expression and epigenetic data are often deposited at NCBI GEO (Gene Expression Omnibus) or ArrayExpress databases. The NCBI SRA (Sequence Read Archive) is a central location for storing sequencing data. A software package called SRA Toolkit provides convenient solutions for downloading large files of sequencing data.


**Figure 2**

A screenshot of a spreadsheet shown in The Cancer Genome Atlas (TCGA) data portal web page when querying available data from breast cancer patients. Each column of this spreadsheet corresponds to a combination of data types and platforms. For each platform, there could be data from levels 1, 2, and/or 3. Usually raw data belong to level 1, and processed data belong to level 2 or 3. One type of genomic data may be collected on multiple platforms. For example, gene expression is measured by both microarray and RNA sequencing (RNA-seq), and two types of DNA methylation arrays (JHU-USC HumanMethylation27 and JHU-USC HumanMethylation450) have been used. A platform name often starts with the institute that processes those tumor samples using that particular platform. For example, BI Genome_Wide_SNP_6 is an Affymetrix 6.0 array from Broad Institute. Each row of this spreadsheet corresponds to a patient. The meaning of the patient ID can be found at **https://wiki.nci.nih.gov/display/TCGA/TCGA+Barcode**. Other terms: BCGSC, Michael Smith Genome Sciences Centre; biotab, a file containing TCGA clinical data; CNV, copy number variation; DNASeq, DNA sequencing; Exp-Gene, gene expression; Exp-Protein, protein expression; HMS, Harvard Medical School; JHU, Johns Hopkins University; MDA, MD Anderson Cancer Center; miRNAseq, microRNA sequencing; SNP, single nucleotide polymorphism; UNC, University of North Carolina; WUSM, Washington University School of Medicine.

There are also growing data resources from large consortium projects. A widely cited example is TCGA. The TCGA data portal allows users to directly download open-access data including de-identified data of clinical and demographic features, mRNA or miRNA expression, copy number alterations, DNA methylation, and protein or phosphoprotein abundance (**Figure 2**). The primary sequence data and genotype data are controlled-access data that can be downloaded from CGHub (Cancer Genomics Hub). The International Cancer Genome Consortium (ICGC) is another large consortium that also collects genomic data from different types of cancers. A few other notable genomic data resources include the Roadmap Epigenomics Project, which focuses on genome-wide epigenetic marks; the Genotype-Tissue Expression project (GTEx), which produces RNA-seq data from different human tissues, and the ENCODE (Encyclopedia of DNA Elements) project, which aims to study all functional elements in the human genome sequence.

Although many datasets are freely available for academic use, becoming familiar with different data resources and making correct use of them requires effort. Many datasets are publicly available but are not well-annotated, making them difficult to use. Databases with standardized uploading protocols are typically easier to use. For example, GEO adopts the MIAME (minimum information about a microarray experiment) standard and has volunteer personnel to constantly check data

**Table 2 Annotation databases**

| Category | Database | URL |
|---|---|---|
| Genome browser | Ensembl | **http://www.ensembl.org/index.html** |
| | UCSC genome browser | **http://genome.ucsc.edu** |
| SNP/indels | dbSNP | **http://www.ncbi.nlm.nih.gov/SNP** |
| Gene structure | GENCODE | **http://www.gencodegenes.org** |
| | Ensembl's Genebuild | **http://www.ensembl.org/index.html** |
| Functional annotation | Pathway Commons | **http://www.pathwaycommons.org** |
| | BioGRID | **http://www.thebiogrid.org** |
| | KEGG | **http://www.genome.jp/kegg** |
| | Gene Ontology (GO) | **http://geneontology.org** |

Terms: BioGRID, Biological General Repository for Interaction Datasets; dbSNP, Single Nucleotide Polymorphism Database; GENCODE, gene features of ENCODE (Encyclopedia of DNA Elements); KEGG, Kyoto Encyclopedia of Genes and Genomes; SNP, single nucleotide polymorphism; UCSC, University of California, Santa Cruz.



quality when new datasets are uploaded. Furthermore, sequencing or genotyping data of human samples often involves issues regarding privacy and legal consent, and thus their datasets need protection through protocols such as dbGAP. The administrative burden to access such data is usually nonnegligible and should be considered before using these datasets.

## 1.3. Genomic Annotation Databases

Genomic annotations, such as locations and functions of genomic features, are valuable knowledge to assist the analysis in any genomic study. Owing to space limitations, we only provide a brief review of selected annotation databases (**Table 2**). Arguably, the most important annotation for most genomic studies is the reference genome. The most recent release of human reference genome is GRCh38.p2 (released on December 8, 2014 by the Genome Reference Consortium), which is the second patch release for the GRCh38 reference assembly. Reference genomes can be accessed online at Ensembl or the UCSC Genome Browser, among other locations. At the DNA level, NCBI dbSNP (Single Nucleotide Polymorphism Database) provides a comprehensive annotation of known SNPs, taken from various sequencing/genotyping projects such as the 1000 Genomes Project. The current version of dbSNP, Human build 142, has a total of 112 million reference SNPs.

Gene structure annotation includes the location of a gene and its exons, as well as its transcripts (i.e., RNA isoforms). Ensembl's Genebuild pipeline automatically annotates genes based on existing evidence of mRNA and proteins in public scientific databases (Curwen et al. 2004). The GENCODE annotation combines the automatic annotation from Ensembl and manual annotation from the Human and Vertebrate Analysis and Annotation team (Harrow et al. 2012). The functional annotation of each gene is incorporated by many gene-centered databases such as NCBI's Entrez Gene database. The Gene Ontology database provides standardized ontology terms for gene functions in three categories: biological process, molecular function, and cellular component (Ashburner et al. 2000). There are also many databases for pathway annotations, such as KEGG (the Kyoto Encyclopedia of Genes and Genomes) and the National Cancer Institute Pathway Interaction Database. Pathway Commons provides a centralized location to store pathway information from multiple databases. Many annotation databases can be conveniently found in the annotation category of Bioconductor, a comprehensive collection of bioinformatics tools built

**1000 Genomes Project:** an international collaboration to produce an extensive public catalog of human genetic variation

on the R language platform. There are also numerous useful databases to systematically catalog existing biological findings such as the GWAS Catalog (disease association findings), COSMIC (catalog of somatic mutations in cancer), miRanda (miRNA target genes), Genomics of Drug Sensitivity in Cancer (drug response in cancer), MIPS (Munich Information Center for Protein Sequences), and Transfac (transcription factor binding motifs), just to name a few.

## 2. HORIZONTAL DATA INTEGRATION

Applications and development of meta-analysis methods to increase statistical power and achieve a consensus conclusion have significantly grown and evolved in the past decade due to the rapid growth of GWAS, gene expression and methylation studies, and the often limited sample size in each study. The ultimate goal is usually to improve detection of differentially expressed genes, disease-associated SNPs, or differentially methylated sites. Due to the large-$p$-small-$n$ nature of omics datasets (and also partly because Microsoft Excel could only accommodate a maximum of 256 columns in the 1990s), samples are usually arranged on the columns, and gene features (SNPs, gene symbols or methylation sites) are on the rows, which reverses the convention of general statistical practices. As a result, when multiple GWAS or transcriptomic studies are combined for meta-analysis, the datasets are laid out horizontally with gene features matched on the rows. This method of multistudy data integration (often called horizontal genomic meta-analysis) is the focus of this section. In contrast, when multiple omics datasets of the same cohort of samples are combined, the datasets are aligned vertically with samples matched on the columns, and this type of data integration is called vertical genomic integrative analysis (discussed in Section 3). For horizontal meta-analysis, interested readers may refer to the following types of publications for details: GWAS meta-analysis review articles (Thompson et al. 2011, Begum et al. 2012, Evangelou & Ioannidis 2013), microarray meta-analysis review articles (Ramasamy et al. 2008, Tseng et al. 2012), and relevant comparative studies (Chang et al. 2013, Wang et al. 2013). Below, we mainly focus on GWAS and transcriptomic meta-analysis to illustrate the basic principles and common issues, as well as to discuss related challenges and opportunities.

### 2.1. Data Collection and Preprocessing

Researchers first determine systematic search and inclusion/exclusion criteria to identify, extract, annotate, and prepare datasets for meta-analysis. This process may involve special data management consideration and tedious preprocessing protocol. In GWAS meta-analysis, for example, raw genotyping data are usually not allowed to be shared without patient consent. The GWAS meta-analysis consortium usually needs to develop a rigorous data exchange protocol to determine sharing of clinical information and summary statistics (e.g., effect size and its standard deviation) for millions of SNPs for meta-analysis. For transcriptomic meta-analysis, determining whether studies have similar underlying biological comparison suitable for meta-analysis is critical. After data preparation, methods may be applied to ensure quality control for meta-analysis (Kang et al. 2012).

### 2.2. Statistical Methods for Meta-Analysis

Many traditional meta-analysis methods have been applied to genomic applications. These include two major categories: combined $p$-values and combined effect sizes. In the first category, Fisher's method and Stouffer's method are probably the most popular. Methods taking the minimum and maximum $p$-values have been used. In the second category, fixed, random or mixed effects models

are popular. In transcriptomic meta-analysis, nonparametric methods based on ranks have also been developed (Hong et al. 2006).

## 2.3. Targeted Biological Objectives and Underlying Hypothesis Settings

An important prerequisite decision behind genomic meta-analysis is to determine the targeted biological objective and the corresponding hypothesis setting. Tseng et al. (2012) demonstrated two hypothesis settings ($HS_A$ and $HS_B$) to detect biomarkers differentially expressed (or SNPs associated to disease) in "all studies" or "one or more studies," respectively. Although $HS_A$ is more often the desired biological objective, $HS_B$ can be considered when study heterogeneity is expected and of research interest (e.g., when studies utilize different tissues; see Section 2.4). These two hypothesis settings are closely related to the traditional union-intersection test and intersection-union test (IUT), and choosing a hypothesis setting affects the selection of a suitable meta-analysis method. For example, Fisher's method combines *p*-values by summation of log-transformed *p*-values. One sufficiently small *p*-value is enough to generate statistical significance and thus, the method is useful for testing $HS_B$ (IUT). Chang et al. (2013) present a comprehensive comparative study to compare different methods for transcriptomic meta-analysis according to different hypothesis settings. Song & Tseng (2014) discuss a robust hypothesis setting to relax $HS_A$ from the stringent requirement of differential expression in "all studies" to "most studies" and propose a solution by employing order statistics.

## 2.4. Cross-Study Heterogeneity

Genomic studies are often heterogeneous across studies owing to the use of different cohorts, experimental protocols, platforms, or tissues used to generate the data. Although the main purpose of meta-analysis is to combine consensus information to improve statistical power, the heterogeneities across studies are also often important. For example, if different tissues are used in different transcriptomic studies, tissue-specific biomarkers are expected and are of concern. In the $HS_B$ hypothesis setting described above, an adaptively weighted concept (i.e., a subset-based approach) and a meta-lasso approach have recently been developed to identify gene-specific subsets of studies that contain differential expression (Li & Tseng 2011, Li et al. 2014) or disease association information (Bhattacharjee et al. 2012, Han & Eskin 2012). The results of these two approaches characterize both homogeneous and heterogeneous signals across studies. In addition to tissue-specific heterogeneity, another common scenario happens in the population structure correction for GWAS meta-analysis (Pritchard et al. 2000). Failure to account for population structure can result in the identification of a spurious association owing to the underlying structure of the population rather than a disease-associated locus (Price et al. 2006).

## 2.5. Horizontal Meta-Analysis for Purposes Other Than Biomarker Detection

Our discussion so far has focused on meta-analysis of multiple genomic datasets to improve biomarker detection. Beyond biomarker detection, the concept of horizontal meta-analysis can be extended to virtually any statistical learning area that has been developed and applied to a single high-throughput experimental dataset. In transcriptomic analysis, for example, gene set analysis (also called pathway analysis) is a popular and powerful tool to characterize biological pathways associated with a disease or condition (Khatri et al. 2012, Newton & Wang 2015). This method can be extended to a meta-analytic setting to improve statistical power and to reach a conclusion with a higher degree of consensus (Shen & Tseng 2010). Other statistical learning areas such as

dimension reduction, clustering (Huo et al. 2016), classification (Kim et al. 2016), and network analysis may also take advantage of the ability to combine information from multiple studies to improve performance; many challenging opportunities remain open in this field.

## 3. VERTICAL DATA INTEGRATION

Vertical integration concerns the analysis of multiple types of data from the same set of samples. Following the fundamental principle of systems biology that biological mechanisms are built upon multiple molecular phenomena acting at different levels, we aim to understand complex phenotypic traits by jointly analyzing different layers of genomic information. Vertical integration tasks are directly linked to the type of biological questions that are posed, and the methods used are correspondingly extremely varied. To determine the appropriate vertical integration approach, one must consider whether the biological question is focused on predictive, regressive (supervised), or exploratory (unsupervised) aims, and how prior biological information is utilized within the statistical analysis. Most approaches to vertical integration are model-based, and we focus our review on these. We discuss some non-model-based approaches, such as pathway recognition and gene set scoring, when we describe examples of analysis strategies aimed at answering specific biological questions (see Section 3.4).

### 3.1. Integrative Clustering

Many diseases exhibit substantial heterogeneity with respect to biological characteristics and clinical outcomes. Motivated by the known influence of genetic aberrations (germline and somatic) and the accessibility of tumor samples, early work focused almost exclusively on cancer and on using simple hierarchical clustering of gene expression profiles to uncover cancer subtypes. The landmark papers by Golub et al. (1999) on leukemia and by Perou et al. (2000) on breast cancer were followed by numerous studies endeavoring to describe molecular subtypes for a large variety of cancers, along with only a few studies related to noncancer pathologies, such as myopathies (Greenberg et al. 2002) or autoimmune diseases (Lee et al. 2011). Cancer genomes exhibit considerable heterogeneity, with abnormalities occurring in different genes among different individuals, posing a great challenge to identify those genes with functional importance and therapeutic implications. It became apparent that to go beyond a straightforward catalog and provide deeper biological insight and clinical significance, additional biological information should be incorporated into the clustering process, and new robust clustering strategies that would simultaneously integrate diverse genomic characteristics were warranted.

**3.1.1. iCluster.** An important step in this direction was made by Shen et al. (2009) with iCluster, an integrative clustering approach to infer latent subtypes based on multiple genomic data types measured on the same samples. They achieve this by specifying a latent model for each data type $\mathbf{X}_j$ (where each dataset is row centered):

$$\mathbf{X}_j = \mathbf{W}_j\mathbf{Z} + \epsilon_j, \qquad j = 1, \ldots, m, \tag{1}$$

where $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_{N-1})$ are the latent subtypes common to the $m$ data types, and $\mathbf{W}_j$ are the coefficient matrices projecting on the latent subspaces. The independent error terms $(\epsilon_1, \ldots, \epsilon_m)$, each associated with a diagonal covariance matrix, represent the residual variance in each dataset after accounting for the correlation between data types. In order to derive a computationally efficient procedure for evaluating the likelihood in Equation 1, a Gaussian latent variable model representation, based on a continuous parametrization $\mathbf{Z}^*$ of $\mathbf{Z}$, is used. An expectation-maximization

(EM) iterative algorithm is employed to derive the reduced representation $\hat{E}(\mathbf{Z}^*|\mathbf{X})$. Additional lasso-type penalties are imposed on the factor scores $\mathbf{W}$ to obtain a sparse solution and pinpoint important features contributing to the clustering. Finally, the class indicators $\mathbf{Z}$ are recovered using a $K$-means procedure on $\hat{E}(\mathbf{Z}^*|\mathbf{X})$ to derive $N$ clusters. Model choice for the lasso penalty and the number of clusters is based on an empirical separability criterion.

This approach, preceded by many filtering steps, was used in the high-profile METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) paper (Curtis et al. 2012) to derive a novel classification of breast cancer patients into clinically meaningful subgroups. Information from inherited variants (CNVs and SNPs) and acquired SCNAs was integrated to define these subgroups. Whereas iCluster was originally formulated for clustering continuous data types, iCluster+ (Mo et al. 2013) extends the framework to cope with both discrete and continuous data, replacing the linear formulation in Equation 1 by a generalized linear one.



**3.1.2. Bayesian integrative clustering approaches.** The METABRIC paper and its potential importance for clinical management of breast cancer opened the door to numerous studies of tumor heterogeneity. It also stimulated the development of alternative clustering approaches that aimed to exploit the power of Bayesian mixture models to increase the flexibility of integrative clustering. Important features of these approaches include the ability to use different types of data (discrete and continuous), the inclusion of a natural assessment of uncertainty provided by the use of Bayesian Dirichlet multinomial models as underlying structure, and the possibility of not assuming the same clustering on all data types. Instead, the Bayesian formulations aim to find related clustering structures across the data types. Two main approaches to modeling cluster dependence have been taken, consisting either of relating clusters or of uncovering common and specific cluster patterns between the data types.

Building on the work of Savage et al. (2010) and Yuan et al. (2011) for integrative clustering of two data types, Kirk et al. (2012) propose a method called multiple dataset integration (MDI). They denote the observed data for gene $i$ in data type $k$ by $X_{ik}$, where $i = 1, \ldots, n$ and $k = 1, \ldots, K$, and specify a Dirichlet-multinomial allocation model for each data type:

- Each gene $i$ is classified into one of $N$ components ($N$ is fixed and the same for each dataset, but components may be empty) with allocation probabilities given by $P(z_{ik} = j) = \pi_{jk}$ for $j = 1, \ldots, N$.
- In each dataset $k$, a mixture model is specified using appropriate parametric densities $f_k$, involving parameters $\Theta_k$.
- Association parameters $\phi_{km} \geq 0$ are introduced to control the strength of association between pairs $(k, m)$ of datasets:

$$P(z_{i1}, \ldots, z_{iK}) \propto \prod_{k=1}^{K} \pi_{z_{ik}k} \prod_{k=1}^{K-1} \prod_{m=k+1}^{K} (1 + \phi_{km} \mathrm{I}_{[z_{ik}=z_{im}]}),$$

where $\pi_{z_{ik}k}$ is the allocation probability of gene $i$ to the component $z_{ik}$ in data type $k$.

Estimation of all the parameters proceeds by stochastic simulation using Gibbs sampling, exploiting natural conjugacy in the model formulation. As clusters are allowed to be empty, $N$ should be sufficiently large, with $N = n/2$ a practical recommended choice. If $\phi_{km}$ is large, then groups of co-clustering genes in dataset $k$ will be encouraged to have the same label in dataset $m$. Interpretation of these parameters and the associated posterior probabilities for a sample $i$ to be fused across the datasets (i.e., for it to have the same label in a subset of data types) allow a rich interpretation of the posterior output.

Lock & Dunson (2013) propose Bayesian consensus clustering (BCC), which aims to simultaneously uncover source-specific clusters for each data type and a common clustering pattern for all data types. Such a decomposition, in line with a tradition of hierarchical modeling of several data sources in epidemiology into common and specific patterns [e.g., Knorr-Held & Best (2001), Ancelet et al. (2012)], makes stronger structural assumptions than MDI. As in MDI, BCC uses a fixed number of clusters $N$ for each data type. BCC considers an overall consensus clustering $C$, with corresponding latent allocation $z_i$, and it links the cluster labels $z_{ik}$ in the different data types to the consensus clustering $C$ through a dependence function $\nu$,

$$P(z_{ik} = j \mid z_i) = \nu(j, z_i, \alpha_k) = \begin{cases} \alpha_k & \text{if } z_i = z_{ik} \\ (1-\alpha_k)/(N-1) & \text{otherwise,} \end{cases} \quad (1)$$

where $\alpha_k \in [1/N, 1]$ controls the level of adherence of data type $k$ to overall clustering and the function $\nu$ aligns the cluster labels in the different data types. As in MDI, estimation of BCC is implemented through Gibbs sampling. Because the BCC algorithm uses a formulation and parametrization that increase linearly with the number of clusters rather than in a quadratic fashion as in MDI, BCC is more scalable to a large number of data sources and samples. However, the appropriateness of the basic assumption that a consensus clustering exists must be evaluated for each case study. Both MDI and BCC use datasets from TCGA to illustrate the performance and interpretability of the clustering patterns uncovered, integrating up to four TCGA data sources each.

It is clear that flexible clustering approaches, which exploit several genomic levels simultaneously, play an important role in integrative genomics. A natural extension of such approaches would be to incorporate additional outcome data (i.e., to use a joint model of features and response in a semi-supervised manner rather than to proceed sequentially with clustering first and then link clusters with survival outcome), as is presented in the METABRIC paper. In the genetic epidemiology context, Papathomas et al. (2012) use a joint clustering of genes and lung cancer outcomes to explore the potential for gene-gene interactions. They adopt a nonparametric Bayesian approach referred to as profile regression (Molitor et al. 2010) that also allows the selection of the important features that drive the clustering. Integration of additional structure in the data, such as spatial organization, into the formulation of integrative clustering models would also be of great interest, as it may provide additional interpretability for the clusters [see Pettit et al. (2014)]. Such extension would be particularly relevant in view of the recent developments in single-cell technologies.

## 3.2. Integrative Regression

Integrative clustering addresses vertical integration for unsupervised tasks. For supervised problems, researchers universally employ regression approaches. When regression tasks involve many more features than samples, the so-called large-$p$, small-$n$ paradigm, one needs additional structure or constraints to derive useful solutions. A large body of work has been developed related to penalized regressions, which produce shrinkage estimates of the regression coefficients, the most common being $\ell_2$ or $\ell_1$ penalties corresponding to ridge or lasso regressions, respectively. Penalized regression approaches for high-dimensional data and their numerous extensions have been thoroughly reviewed by Bühlmann et al. (2014). In this section, we focus on Bayesian approaches, which combine variable selection for high-dimensional problems with integrative genomics tasks.

**3.2.1. Including prior information into variable selection.** Given a set of $p$ covariates $\{X_j, j = 1, \ldots, p\}$, let us consider the regression model of outcome $y$ on the set of covariates

$$y = \mathbf{X}\boldsymbol{\beta} + \epsilon, \quad \epsilon \sim N\left(0, \sigma_\epsilon^2 I_n\right), \tag{2}$$

where $y = (y_1, \ldots, y_n)^T$, $\mathbf{X}_{n\times p}$ is the covariate matrix, $n$ is the number of samples with $p \gg n$, and $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^T$ is the vector of regression coefficients (response and covariates are assumed centered).

Bayesian variable selection methods typically include binary variable selection indicators $\gamma_j$ indicating if variable $X_j$ is included or not ($\gamma_j = 1$ if $\beta_j \neq 0$ and $\gamma_j = 0$ if $\beta_j = 0$) and aim to explore the vast set of $2^p$ possible models corresponding to $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$. Alternatively, spike and slab priors for the regression coefficients have also been considered (Ishwaran & Rao 2005). Focusing for now on conditional formulations, Equation 2 reduces to

$$y \mid \boldsymbol{\gamma} = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \epsilon, \quad \epsilon \sim N\left(0, \sigma_\epsilon^2 I_n\right), \tag{3}$$

where $\boldsymbol{\beta}_\gamma$ is the vector of nonzero coefficients, $\mathbf{X}_\gamma$ is the $n \times p_\gamma$ reduced design matrix with columns corresponding to $\gamma_j = 1$, and $p_\gamma$ is the overall number of nonzero coefficients.

Full Bayesian inference requires prior specification for the regression coefficients. In order to explore the vast model space efficiently, conjugate priors for the regression coefficients $\boldsymbol{\beta}_\gamma$ are commonly adopted so that the regression coefficients can be integrated out. Both independent priors (Hans et al. 2007) and Zellner's $g$-prior structure with a hyperprior on $g$ have been adopted (Bottolo & Richardson 2010). A range of efficient stochastic algorithms to explore the vast model space have been proposed, including the stochastic shotgun sampler (Hans et al. 2007) and others inspired by population Monte Carlo techniques, such as the evolutionary stochastic search (ESS) sampler (Bottolo & Richardson 2010, Bottolo et al. 2011a).

The prior model of the binary indicators has a direct influence on the sparsity of the model space. On the one hand, exchangeability might be assumed. Sparsity can then be tuned to encompass prior assumptions on the mean and variability of the overall expected number of selected covariates (Bottolo & Richardson 2010). On the other hand, specific external information $W_j$ might be available on each of the covariates $X_j$, information that could make the selection of $X_j$ more or less likely. Such a situation arises, for example, in genetic association studies in which additional functional characterizations of the SNPs in terms of genomic regions or functional annotation might be relevant. To integrate such information in a flexible manner, a natural extension of Equation 3 is to specify a hierarchical model for $\{\gamma_j\}$ and use a probit link for linking the underlying probabilities to the external information:

$$\gamma_j \sim B(\pi_j), \quad \pi_j|\alpha \sim \Phi(\alpha_0 + \mathbf{W}_j'\alpha_1), \; j = 1, \ldots, p. \tag{4}$$

If $\alpha_1 \simeq 0$ then the model in Equation 4 is equivalent to a standard exchangeable prior on the selection indicators. Estimating $(\alpha_0, \alpha_1)$ together with Equation 3 allows us to quantify the influence of the external information. Quintana & Conti (2013) propose such an extension and illustrate its benefits in a genetic association study of smoking cessation involving 121 SNPs. For each SNP, they integrate external information on gene regions and on a quantitative association with a nicotine metabolite ratio. Integrative regression can also be used when building directed networks, as discussed in Section 3.3.

In addition to quantitative information, structural and distributional information can also be integrated in a variable selection framework to improve inference. For example, Stingo et al. (2011) include prior information on gene networks to better select discriminatory variables. They model

the joint distribution of the binary selection indicators $\{\gamma_j\}$ as a Markov random field

$$P(\gamma_j \mid d, f, k \in N_j) = \frac{\exp\left(\gamma_j(d + f\sum_{k\in N_j}\gamma_k)\right)}{1 + \exp\left(\gamma_j(d + f\sum_{k\in N_j}\gamma_k)\right)},$$

where $N_j$ is the set of direct neighbors of variable $j$ in a preset graph (e.g., extracted from the KEGG database), $d$ controls the sparsity of the model, and $f$ controls the strength of spatial structure; both $d$ and $f$ are fixed.

The Bayesian approaches discussed above for integrating information into the variable selection process have analogs in the penalized regression context. Group lasso (Yuan & Lin 2006) allows groups or networks of genes to be viewed as additional information to tailor the penalization. Pan et al. (2010) refine these ideas in the genomic context by using a group penalty based on KEGG pathway information to predict survival of glioblastoma patients. To incorporate external information provided by additional sources of data, Bergersen et al. (2011) propose a weighted lasso approach in which the weights are inversely proportional to a quantitative function linking external information, response, and covariates. An additional tuning parameter controlling the relative strength of all the weights is calibrated through cross-validation. This flexible approach allows a variety of external information to be straightforwardly incorporated into the analysis (e.g., copy number alterations when looking for prognostic gene expression signatures) and was shown to improve predictive ability.

**3.2.2. Multiple response model.** When faced with a set of correlated responses or related phenotypes, one may perform joint regression analysis of these responses as another way of borrowing information to increase sensitivity. Multiple response models extend the single outcome regression (Equation 3) to multidimensional responses $\mathbf{Y}(n \times \ell)$, where $\ell$ is the number of responses, by considering the residual variance-covariance between the responses $\Sigma$ $(\ell \times \ell)$. The likelihood (Equation 2) becomes

$$\mathbf{Y}|\boldsymbol{\gamma} = \mathbf{X}_{\boldsymbol{\gamma}}\mathbf{B}_{\boldsymbol{\gamma}} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(I_n, \Sigma), \tag{5}$$

where $\mathbf{X}_{\boldsymbol{\gamma}}(n \times p_\gamma)$ as above, and $\mathbf{B}_{\boldsymbol{\gamma}}(p_\gamma \times \ell)$ now represents the matrix of regression coefficients for the selected variables. Equation 5 assumes that the predictors have an effect (possibly different) on multiple outcomes at once, and the borrowing of information is effected through the correlation between the responses. The $2^p$ model search task is similar to the one discussed for single response (see Section 3.2.1), and Bayesian algorithms can be extended straightforwardly to multiple response models.

This approach was used by Bottolo et al. (2013) to analyze groups of correlated lipid phenotypes. They were inspired by the known structure of HDL (high-density lipoprotein) and LDL (low-density lipoprotein) cholesterol pathways and analyzed different combinations of lipid biomarkers (triglyceride, HDL and LDL cholesterol, apolipoprotein A1, and apolipoprotein B) following Equation 5; these were then regressed on a genome-wide set of 273,675 SNPs derived from the Affymetrix Genome-Wide Human SNP array 6.0 (tagged $r^2 > 0.8$). To cope with the challenging computational task of performing model exploration on this large set of SNPs observed on 3175 individuals, Bottolo et al. (2013) developed a version of the ESS algorithm based on graphics processing units (GPU). They derived synthetic measures of evidence such as a list of top models, together with estimates of their posterior probability and Bayes factors against the null model, as well as marginal posterior probability of inclusion for each SNP (using model averaging) rescaled to be comparable across different combinations. The results provide new insight into the genetic control of lipid pathways, refining some of the previous GWAS results.

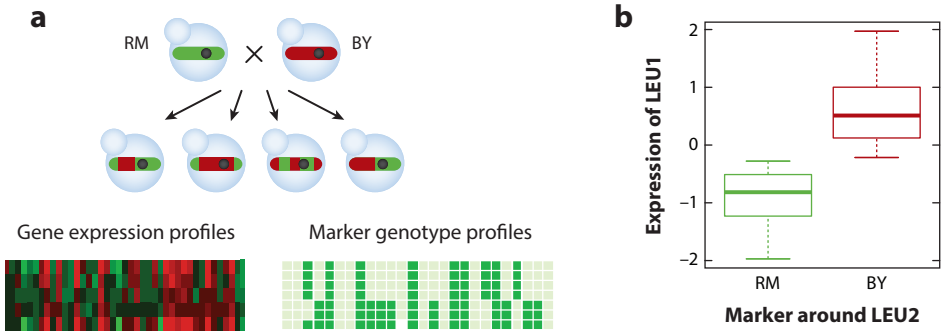



**Figure 3**

(*a*) An illustration of the first genome-wide gene expression quantitative trait loci study (Brem et al. 2002) conducted in yeast sergeants (offspring) from a cross of two yeast strains, denoted by RM (*green*) and BY (*red*). Because yeast has a haploid genome, the genotype data in these yeast sergeants are binary. (*b*) The association between the expression of one gene LEU1 and the genotype of one single nucleotide polymorphism around gene LEU2.



**3.2.3. Joint regression of two omics datasets and gene expression quantitative trait loci models.** Many biological questions can be expressed under the generic framework of performing the joint regression analysis of two or more different types of genomics datasets. In this section, we focus on analyses in which a large number of responses are regressed on a very large number of predictors. The multiple response models described in Section 3.2.2, which assume that a set of predictors affect all the responses at once, are not well adapted to analyses involving a large number of responses. A canonical example of genomic studies involving a large number of responses are the so-called gene expression quantitative trait loci (eQTL) studies, which investigate the genetic control of expression by regressing expression profiles on DNA variants (**Figure 3**). Other examples are mQTL (metabolite quantitative trait loci) studies, which link DNA variations to metabolite synthesis (Marttinen et al. 2014), and studies investigating the influence of SCNAs on tumor gene expression. Flexible ways of borrowing information between the high-dimensional phenotypes are required to increase power. In other words, rather than testing the association between each pair (marker × expression) separately and subsequently facing a huge multiplicity adjustment, the high-dimensional set of gene-expression responses are treated as related outcomes. The statistical aims are thus expanded not only to uncover the multivariate association of each (expression) response with a large number of features (e.g., genetic markers) but also to highlight the features that are associated with many responses. Finding key control points associated with the expression of many genes, sometimes called hot spots, is an important step toward a better understanding of biological pathways.

Different approaches have been proposed for discovering regression links between a large number $q$ of responses ($\boldsymbol{y}_k$, $1 \leq k \leq q$) and a large set of predictors $\boldsymbol{X}$ in a way that exploits the relatedness of the responses. Penalized approaches use structured regularization to account for the correlation of the responses. For example, Peng et al. (2008) propose a combination of $\ell_1$ and $\ell_2$ penalties to encourage the detection of master regulators, whereas Kim & Xing (2012) use a tree-guided lasso to account for the relationship between the genes. An early Bayesian approach is the mixture over markers method (Kendziorski et al. 2006), which associates each response with any of the $p$ predictors (or none of them) via a mixture model, so each response is associated with at most one marker, a workable but restrictive assumption. Stochastic partition approaches, in which the responses are partitioned into disjoint subsets that have a similar dependence on a

subset of covariates, have also been implemented in eQTL analyses under strong assumptions on the commonality of effects within the blocks (Monni & Tadesse 2009; see also Zhang et al. 2010 for an application of Bayesian partitioning to find pleiotropic and epistatic eQTL modules).

Bayesian approaches combining high-dimensional variable selection for each response with a hierarchical structure on the selection indicators have the benefit of being fully multivariate while retaining scalability. The key quantities that are involved in such models are as follows:

- the latent binary vectors $\boldsymbol{\gamma}_k = \left(\gamma_{k1}, \ldots, \gamma_{kj}, \ldots, \gamma_{kp}\right)^T$ for each regression of $\boldsymbol{y}_k$ on $\boldsymbol{X}$, where each indicator has a Bernoulli prior $p\left(\gamma_{kj} \left| \omega_{kj} \right.\right) = \omega_{kj}^{\gamma_{kj}} \left(1 - \omega_{kj}\right)^{1-\gamma_{kj}}$;
- $\boldsymbol{\Gamma} = (\gamma_{kj}, 1 \leq k \leq q, 1 \leq j \leq p)$, the $(q \times p)$ matrix of selection indicators; and
- a hierarchical structure for the matrix of prior probabilities for $\boldsymbol{\Gamma}$

$$\boldsymbol{\Omega} = \begin{bmatrix} \omega_{11} & \cdots & \omega_{1j} & \cdots & \omega_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \omega_{k1} & \cdots & \omega_{kj} & \cdots & \omega_{kp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \omega_{q1} & \cdots & \omega_{qj} & \cdots & \omega_{qp} \end{bmatrix}$$

  that facilitates sparsity control in each regression as well as information borrowing across responses to highlight important predictors common to several responses.

Different prior structures for $\omega_{kj}$ have been proposed. Bottolo et al. (2011b) introduce a multiplicative parametrization $\omega_{kj} = \omega_k \times \rho_j$ with $\omega_k \sim \text{Beta}\left(a_{\omega_k}, b_{\omega_k}\right)$, $\rho_j \sim \text{Gam}\left(c_{\rho_j}, d_{\rho_j}\right)$, subject to the constraint that $0 \leq \omega_{kj} \leq 1$. In this choice of parametrization, $\rho_j$ captures the propensity for predictor $j$ to influence several outcomes at the same time, and $\omega_k$ controls the sparsity of each regression. Scott-Boyer et al. (2011) propose a mixture model for $\omega_{kj}$ with an atom at zero to reduce the false discovery rate, a Beta distribution for the second mixture component, and a SNP-specific mixture weight: $\omega_{kj} = p_j \delta_0(\omega_{kj}) + (1 - p_j)\text{Beta}(a_j, b_j)(\omega_{kj})$. Both approaches are implemented by Markov chain Monte Carlo (MCMC). The choice of a $g$-prior structure for the regression coefficients in Bottolo et al. (2011b) allows the regression coefficients to be fully integrated out and to use a hierarchical extension of their ESS sampler to traverse the model space, whereas the prior structure defined by Scott-Boyer et al. and implemented in their algorithm iBMQ (integrated Bayesian modeling of eQTL data; Imholte et al. 2013) requires joint updating of the variable selection and regression coefficients. Despite efficient MCMC implementation, fully Bayesian joint eQTL analysis strategies are nevertheless quite demanding in terms of computational time and would typically need to be run in parallel on each chromosome on only a few thousand genes. Ultrafast implementation of a linear model, which tests the association of each SNP with each transcript and was implemented by Shabalin (2012), could be used as a preselection step.

Both Bottolo et al. (2011b) and Scott-Boyer et al. (2011) make the simplifying assumption of no residual dependence between the responses conditional on the selected model. Adding a model of the residual structure to the previous setup, in a framework akin to the seemingly unrelated regressions model, has been proposed by Bhadra & Mallick (2013). The additional computational complexity is severe, and such approaches will not easily scale up for large $q$. To encompass additional nuisance correlation in a computationally feasible manner, Stegle et al. (2010) and Fusi et al. (2012) propose a variational Bayesian approach that accounts for additional known and hidden sources of variation using a computationally tractable latent factor approach. The factors can be used as covariates in standard eQTL mapping or can be interpreted as corresponding to transcription factor activations if additional biological information is provided to the model. The efficient software PEER (probabilistic estimation of expression residuals; Stegle et al. 2012),

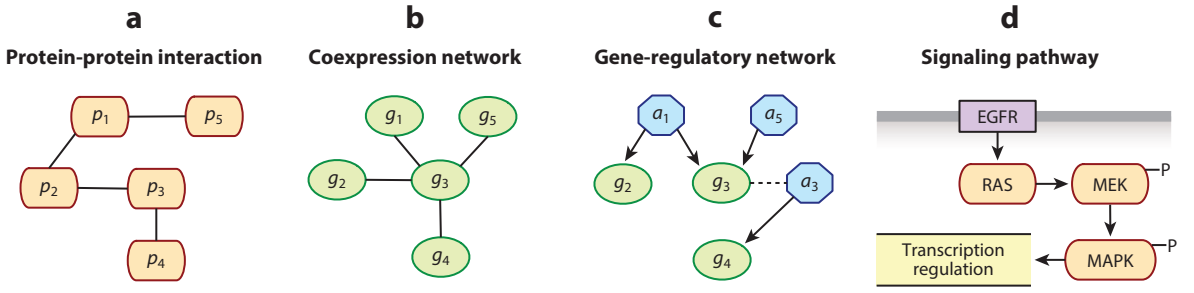



**Figure 4**

Examples of biological networks. (*a*) In protein-protein interaction, the edges are detected experimentally via physical interaction. (*b*) In coexpression networks, the edges are inferred from the expression data of a group of samples as dichotomized correlations or partial correlation matrices. (*c*) Gene-regulatory network, in which the octagons indicate protein activity and the ovals indicates gene expression; the dashed line between $g_3$ and $a_3$ indicates that $a_3$ is the activity of the protein that is encoded by $g_3$. (*d*) In a signaling pathway, EGFR (epidermal growth factor receptor) is a receptor located at the cell surface that can be activated by epidermal growth factor. The symbol $^{-P}$ around a protein indicates a phosphorylation of the corresponding protein. Other abbreviations: *a*, protein activity; *g*, gene expression; MAPK, mitogen-activated protein kinase; *p*, protein abundance.

which is designed to uncover such hidden factors, has been used to reanalyze several eQTL studies that show an increase of power for eQTL detection. In the same spirit, Bayesian reduced rank regression has been used by Marttinen et al. (2014) to analyze gene-metabolome associations, account for known factors, and combine information over multiple SNPs and phenotypes.

## 3.3. Graphical Models

Graphical models, or biological networks, are powerful tools to describe the relationships between different biological entities. Commonly used biological networks include protein-protein interaction networks, coexpression networks, gene-regulatory networks, and signaling pathways (**Figure 4**). Many statistical methods have been developed to construct or exploit the knowledge from such networks to analyze genomic data. We focus on two approaches in this subsection. First, we describe the construction of directed acyclic graphs (DAGs) for genes using gene expression and eQTL data. Second, we describe the inference of miRNA-gene regulation using miRNA expression, gene expression, and annotation data.

**Protein-protein interaction:** physical interaction of two or more proteins for various biological functions, e.g., signal transduction

**Coexpression networks:** networks in which two genes are connected if their expressions are dependent, with or without conditioning on the expression of other genes

**Gene-regulatory networks:** networks in which an edge indicates a transcriptional regulation—regulators can be transcription factors or other molecules such as miRNA

### 3.3.1. Gene expression quantitative trait loci–guided directed acyclic graph construction.

A coexpression network is usually an undirected graph. However, there are many situations in which a directed graph is desirable, for example to infer the consequence of a perturbation by a drug. DAG models have been used to construct directed graphs using gene expression data (Neto et al. 2008, 2010; Hageman et al. 2011; Bühlmann et al. 2014). In such a DAG, each vertex represents a gene and each edge represents a direct causal relation between two genes. For example, an edge $g_1 \rightarrow g_2$ implies that perturbation of $g_1$ alters $g_2$ whereas changes on $g_2$ leave $g_1$ unaffected. It is well known that interventions or perturbations are needed to infer causal relations. However, a huge number of interventions on gene expression (e.g., gene knockouts) are needed to infer causal relations of thousands of genes, and such interventions are not feasible yet. The eQTLs of gene expression provide natural perturbations to the expression of a large number of genes. The passing of DNA alleles to offspring can be considered randomized experiments on DNA genotype (i.e., Mendelian randomization; Smith 2007, Sheehan et al. 2008), and the experimental design (i.e., interventions on DNA genotype rather than gene expression) is also

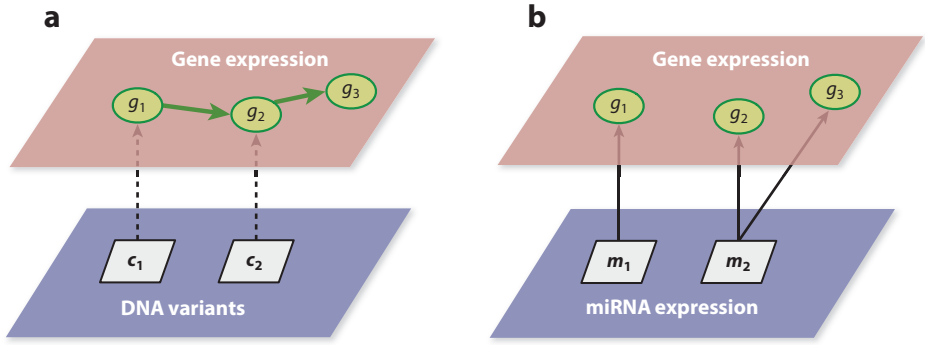



**Figure 5**

Statistical approaches to use graphical modeling in integrative genomics, (*a*) to infer directed coexpression network with the aid of DNA variation information and (*b*) to infer the regulatory relation between miRNA (microRNA) and gene expression. Other terms: *c*, copy number; *g*, gene expression; *m*, miRNA expression.

consistent with our intuition that DNA genotype affects gene expression rather than vice versa (Chen et al. 2007).

The example in **Figure 5*a*** illustrates a situation in which eQTLs can help to estimate edge directions in a DAG. Consider genes $g_2$ and $g_3$, which are coexpressed, and thus there is an undirected edge $g_2 - g_3$ in the graph. Without external data, we cannot distinguish $g_2 \rightarrow g_3$ and $g_2 \leftarrow g_3$, because two DAGs encode the same dependence assumption and have the same likelihood: $L(g_2 \rightarrow g_3) = f(g_3|g_2)f(g_2) = f(g_2, g_3) = f(g_2|g_3)f(g_3) = L(g_3 \rightarrow g_2)$. If $g_2$ has an eQTL, denoted by $c_2$, then the partially directed graph is $c_2 \rightarrow g_2 - g_3$, and the possible DAG is $c_2 \rightarrow g_2 \rightarrow g_3$ or $c_2 \rightarrow g_2 \leftarrow g_3$. These two graphs can be distinguished because they encode different conditional independence assumptions. $c_2 \rightarrow g_2 \rightarrow g_3$ implies $c_2 \perp g_3|g_2$ and $c_2 \rightarrow g_2 \leftarrow g_3$ implies $c_2 \sim g_3|g_2$, and thus these scenarios have different likelihoods. To understand the reason that $c_2 \sim g_3|g_2$, one may consider an example "rain $\rightarrow$ wet grass $\leftarrow$ sprinkler," where given the event that grass is wet, the two parent vertices rain and sprinkler are dependent.

To use eQTLs to derive a causal gene expression network, we also need to separate direct and indirect eQTL effects. Using the example in the previous paragraph and assuming the causal relation is $c_2 \rightarrow g_2 \rightarrow g_3$, $c_2$ may appear to be an eQTL for both $g_2$ and $g_3$. For the purpose of DAG estimation, we need to know that $c_2$ directly affects $g_2$ but indirectly affects $g_3$. Such information can be obtained by separating *cis*-eQTL and *trans*-eQTL using RNA-seq data (Sun 2012, Sun & Hu 2013). All of the *cis*-eQTLs directly influence their target genes, and a *trans*-eQTL may influence its target's expression directly or indirectly. Therefore, it is desirable to use only *cis*-eQTLs for DAG construction.

Neto et al. (2008) develop the quantitative trait loci (QTL) directed dependency graph (QDG) method and implement it in the R package qtlnet. The QDG method was originally designed to study the relations of multiple phenotypes given their QTLs, though it can be applied for eQTL studies as well. The QDG method assumes that multiple QTLs associated with a given set of traits have previously been determined, and it has the following steps. First, construct a DAG skeleton from the PC (Peter-Clark) algorithm, which is a popular algorithm for DAG skeleton construction (Spirtes et al. 2001). Second, distinguish between QTLs with direct and indirect effects. Third, orient each edge by LOD score (the $\log_{10}$ likelihood ratio for the edge $Y_i \rightarrow Y_j$ versus $Y_j \rightarrow Y_i$ given all the vertices—either phenotype or DNA genotype—connected to $Y_i$ or $Y_j$).

**Signaling pathways:** pathways for signal transduction, starting with activation of a receptor by extracellular molecules; the receptor triggers a series of events leading to cell response

***cis*-eQTL:** located on the same chromosome as its target gene; influences the gene expression in an allele-specific manner

***trans*-eQTL:** can be located anywhere in the genome; influences the gene expression of both alleles to the same extent

**DAG skeleton:** an undirected graph constructed by removing the directions of all the edges within a DAG

Fourth, randomly choose an order of all the edges, and following this order, sequentially update the directions of the edges using the LOD score conditioning on the vertices that are parents of $Y_i$ or $Y_j$. Fifth, repeat step four 1,000 times and choose the graph with the highest score as determined by a likelihood-based measure of goodness of fit.

In a later paper, Neto et al. (2010) develop a new method called QTLnet, which jointly estimates the graphical structure of phenotypes and the underlying genetic architecture. This method would be computationally too demanding to study genome-wide eQTL data with tens of thousands of genes and millions of SNPs. In addition, the genetic architecture of human gene expression is relatively simple, and the vast majority of eQTLs are local eQTLs. Therefore, it may be a reasonable approximation to assume that the genetic architecture only involves local eQTLs and then estimate eQTLs before DAG estimation. In contrast to QDG, which reports the most likely graph, QTLnet reports graph structure based on Bayesian model averaging. In other words, the posterior probability of edge $Y_i \rightarrow Y_j$ is the summation of the posterior probabilities of the graphs that have the edge $Y_i \rightarrow Y_j$.

Another type of approach for graphical model estimation is structural equation modeling that permits both cyclic and acyclic graphs. Li et al. (2006) employ a score-based model selection method. Logsdon & Mezey (2010) estimate the skeletons of networks by applying an adaptive lasso regression for each gene expression trait and then transforming the skeleton into a DAG or a directed cyclic graph based on eQTL perturbations. Cai et al. (2013) extended the work of Logsdon & Mezey (2010) by providing the adaptive lasso with a set of initial parameter estimates from penalized regressions using the lasso penalty.

**3.3.2. Construction of microRNA regulation networks.** Recent studies have shown that miRNA, a class of short noncoding RNA molecules (21–24 nucleotides), may play an important role in transcriptional and posttranscriptional regulation of gene expression (Pasquinelli 2012). The human genome may encode more than 1,000 miRNA, which may target more than half of human transcripts. One miRNA sequence may match the complementary sequences of one or more mRNA, and thus the first miRNA can bind these base-paired mRNA sequences, which leads to mRNA degradation or represses the translational process. Therefore, overexpression of miRNA usually reduces the expression of its targets. In plants, an miRNA sequence is often perfectly or almost perfectly matched with its targets. However, animal miRNA sequences typically exhibit only partial complementarity to their mRNA targets. A seed region approximately 6–8 nucleotides long at the 5′ end of an animal miRNA sequence is thought to be an important determinant of target specificity (Pasquinelli 2012). Many computational approaches have been developed to predict miRNA targets based on sequencing similarity. However, these methods have limited accuracy owing to the relatively low target specificity based on sequence data alone. Motivated by this problem, researchers have developed several methods to integrate gene expression, miRNA expression, and miRNA target annotation based on sequence similarity to infer the miRNA regulatory network (Muniategui et al. 2013). Here, we briefly review a Bayesian graphical model approach (Stingo et al. 2010).

Denote the expression of $G$ genes by $\mathbf{Y} = (Y_1, \ldots, Y_G)$, and the expression of $M$ miRNA sequences by $\mathbf{X} = (X_1, \ldots, X_M)$. Stingo et al. (2010) construct a DAG for these $G + M$ variables where the only allowable edges are those of the form of $X_i \rightarrow Y_k$ where $i = 1, \ldots, M$ and $j = 1, \ldots, G$. In other words, they assume that $X_i \perp X_j$ for any $i, j = 1, \ldots, M$ and $i \neq j$, and $Y_k \perp Y_l|\mathbf{X}$ for any $k, l = 1, \ldots, G$ and $k \neq l$ (**Figure 5*b***). Because the marginal distribution of $\mathbf{X}$ does not affect the estimation of regulatory relations, the assumption $X_i \perp X_j$ is a reasonable choice to simplify the computation. When sample size is large enough, one may further relax the assumption $Y_k \perp Y_l|\mathbf{X}$, though imposing this assumption may be the best one can do given

limited sample size and/or computational power. An additional assumption is that all the edges must point from $X_i$ to $Y_j$, which is justifiable because of the regulatory role of miRNA. Given such an underlying DAG, the problem reduces to $G$ regression problems where the $g$th problem aims to select those miRNA sequences that regulate the $g$th gene.

Stingo et al. (2010) assume a linear model with Gaussian errors such as

$$Y_g = -\sum_{m=1}^{M} X_m \beta_{gm} + \epsilon_g, \tag{6}$$

where the $\epsilon_g$ values are independent and normally distributed with a mean of zero and a standard deviation of $\sigma_g$. $\mathcal{N}(0, \sigma_g)$ and the negative sign in front of the term $\sum_{m=1}^{M} X_m \beta_{gm}$ indicate that miRNA repress gene expression. The prior distributions for the regression coefficients $\beta_{gm}$ are

$$\pi(\beta_{gm}|\sigma_g, r_{gm}) = r_{gm}\text{Gam}(1, c\,\sigma_g) + (1 - r_{gm})I_{(\beta_{gm}=0)}, \tag{7}$$

where Gam() indicates a gamma distribution, $I_{(\beta_{gm}=0)}$ is an indicator function, $r_{gm} = 1$ if the $m$th miRNA regulates the $g$th gene, and $r_{gm} = 0$ otherwise. The key part of the method is to incorporate the annotation based on sequencing similarity, and it is implemented as the prior distribution for $r_{gm}$:

$$\log\left(\frac{P(r_{gm} = 1)}{1 - P(r_{gm} = 1)}\right) = \eta + \sum_{u=1}^{U} s^u_{gm}\tau_u, \tag{8}$$

where $s^u_{gm}$ is a score describing the degree of confidence that the $m$th miRNA sequence regulates the $g$th gene. The regression coefficients $\tau_u$ are an additional set of parameters with a hyperprior set as a gamma distribution Gam($a_\tau$, $b_\tau$). Then Stingo et al. (2010) design an MCMC approach to sample all the parameters, of which the $r_{gm}$ values are of primary interest because they indicate whether the $m$th miRNA sequence regulates the $g$th gene.

**3.3.3. Inference of each gene's contribution to the activity of a pathway.** Most human diseases are complex (e.g., diabetes or cancer) and are associated with mutations or perturbations of multiple genes. Such complex diseases may be better described at the pathway level. For example, two patients may have different sets of mutations, but each set may modify the activity of the same pathway. Therefore, to study the importance of a gene that belongs to a known pathway, one may quantify the change of pathway activity by turning this gene on or off. As shown in **Figure 4*c*,*d***, the contribution of a gene to a pathway (e.g., a transcriptional regulation or signaling pathway) should be measured by its protein activity, which could be the abundance of an active form (e.g., a specific phosphoprotein) compared to the abundances of the states of other proteins within the pathway. These quantities are often latent variables that cannot be directly measured on the genome scale using current techniques. PARADIGM (pathway recognition algorithm using data integration on genomic models; Vaske et al. 2010) is a popular computational method that addresses this challenge by integrating different types of genomic data and pathway annotation.

In the PARADIGM model, Vaske et al. (2010) assume the pathway information is known. They consider a pathway as a graph, and a vertex of this graph can be a protein-coding gene, a protein complex, a gene family, an abstract process (e.g., apoptosis), or some other biological entity. Each vertex has three states—activated, nominal, or deactivated, relative to a control level—and is encoded as 1, 0 or −1, respectively. Therefore, the graph is a factor graph, and such a simplified assumption of three states greatly reduces the difficulty of model estimation. Each edge of this graph has a sign that indicates whether the parent vertex has a positive or negative influence

**Protein complex:** a group of two or more proteins that are physically associated; its function may require each individual protein to be active

**Gene family:** a collection of genes in which any single gene is sufficient to perform a specific function

**Apoptosis:** programmed cell death

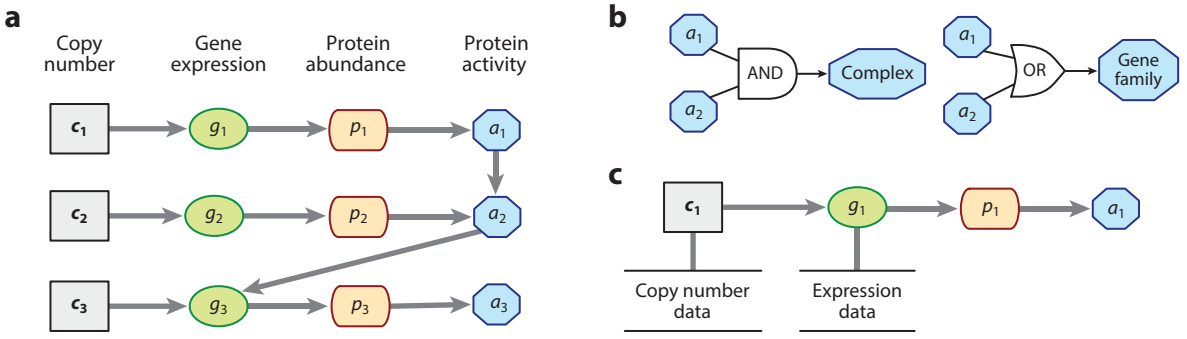



**Figure 6**

The PARADIGM method of integrative genomics (Vaske et al. 2010). (*a*) The graphical model of underlying states, where $a_1 \rightarrow a_2$ indicates that activated protein 1 regulates the activity of protein 2. $a_2 \rightarrow g_3$ indicates that activated protein 2 regulates the expression of $g_3$. (*b*) Examples of AND and OR relations. (*c*) Adding observed data into the graph. Here, observed data, like the underlying state, has three values, −1, 0, or 1. Having observed data in certain vertices means that the underlying states of those vertices are known. Abbreviations: *a*, protein activity; *c*, copy number; *g*, gene expression; *p*, protein abundance.



on the child vertex. PARADIGM includes four entities for the $j$th protein-coding gene: DNA copy number ($c_j$), mRNA expression ($g_j$), protein abundance ($p_j$), and protein activity ($a_j$) (see **Figure 6*a***) for an example of three protein-coding genes). Directed edges with positive signs are introduced as $c_j \rightarrow g_j \rightarrow p_j \rightarrow a_j$. The relations between protein coding genes are introduced based on pathway annotation. For example, if activated protein 1 induces the activity of protein 2, a directed edge $a_1 \rightarrow a_2$ with a positive sign is added. If activated protein 2 represses the expression of gene 3, a directed edge $a_2 \rightarrow g_3$ with a negative sign is added (**Figure 6*a***).

The graph allows one to compute the expected state of the $i$th vertex, denoted by $\mu_i$, given its parents. Assuming the parents contribute additively, $\mu_i = \text{sign}(\sum_{j \in \text{Pa}_i} \mu_j \beta_{ji})$, where $\text{Pa}_i$ denotes the parent set of the $i$th vertex, and $\beta_{ji} = 1$ or $-1$ is the sign of the edge $j \rightarrow i$. PARADIGM also allows the contributions from all the parents to be summarized by an AND or OR operation (**Figure 6*b***). We use $X_i$ to denote the underlying state of the $i$th vertex. If $X_i$ is unobserved, it follows a categorical distribution across the three classes such that $P(X_i = a) = 1 - \epsilon$ if $a = \mu_i$ and $P(X_i = a) = \epsilon/2$ otherwise, where $\epsilon$ is a small value, e.g., $\epsilon = 0.001$. For some of the vertices, the values of $X_i$ are observed (assuming no measurement error), and the purpose of the PARADIGM method is to infer the state of the unobserved $X_i$ values. They employ an EM algorithm to infer such hidden states. Finally, an integrated pathway activity (IPA) is estimated for each biological entity. The term IPA may be misleading: It does not estimate pathway activity per se. Instead, it calculates how much each entity contributes to the pathway activity. Specifically, let $\ell(i, a) = \log[P(D|X_i = a)/P(D|X_i \neq a)]$, which is the log-likelihood ratio comparing the situation $X_i = a$ versus $X_i \neq a$. Then,

$$\text{IPA}(i) = \begin{cases} \ell(i, 1) & \text{if } \text{argmax}_a\, \ell(i, a) = 1 \\ -\ell(i, -1) & \text{if } \text{argmax}_a\, \ell(i, a) = -1 \\ 0 & \text{otherwise.} \end{cases}$$

In other words, IPA($i$) is the signed log-likelihood ratio if the most likely state is 1 or −1, and IPA($i$) is 0 otherwise. Vaske et al. (2010) further demonstrate that clustering IPAs may reveal meaningful clusters that cannot be identified by clustering each type of genomic data directly.

## 3.4. Other Methods for Integrative Genomics

In the previous sections, we have highlighted some recent work in integrative genomics. Because integrative genomics is a very active research area with huge amount of literature, we cannot give an exhaustive review of all work in this area. In the rest of this section, we summarize some methods that aim to answer specific biological questions.

**3.4.1. Phenotype association/prediction.** Xiong et al. (2012) propose a method called gene set association analysis to identify disease-associated gene sets. They first assess the association between a phenotype and the expression and genotype of each gene separately. Then they combine the z-statistics of differential expression and genotype association using Fisher's method for each gene and use the combined gene-specific test statistic for gene set enrichment analysis (Newton & Wang 2015). Instead of analyzing each type of data separately, Tyekucheva et al. (2011) first summarize each type of genomic data at the gene level and then study the association between a gene and a phenotype using one regression model. In this approach, the phenotype is used as the response variable, and different types of genomic data are used as covariates. They score this gene using a test statistic for the null hypothesis that the regression coefficients for all the genomic data are zero. Then this test statistic is used for gene set enrichment analysis.

Another question that is often of interest is whether one type of genomic data mediates the effect of another type of genomic data on phenotype—for example, whether gene expression mediates the effect of DNA genotype on disease outcomes. Huang et al. (2014) use mediation analysis to address this question and quantify a SNP's direct and a genotype's indirect effects (mediated by gene expression) on disease outcomes.

In addition to association testing, genomic data can be used for the prediction of phenotypes. Many genomic features have relatively small effects on the phenotype, so there may not be enough power to identify such genomic features by association testing. However, one may still be able to perform predictions without selecting phenotype-associated genomic features. An example is OmicKriging (Wheeler et al. 2014). Kriging is a well-known geostatistical method for predicting spatially measured outcomes using observations from nearby locations (Cressie 1993). In OmicKriging, Wheeler et al. (2014) assume the similarity matrix of phenotype data across all samples is $\mathbf{\Sigma} = \sum_{j=1}^{J} \theta_j \mathbf{S}_j + (1 - \sum_{j=1}^{J} \theta_j)\mathbf{I}$, where $\mathbf{S}_j$ is the similarity matrix for the $j$th type of genomic data and $\mathbf{I}$ is an identity matrix to capture variance due to environmental factors. Given such phenotype similarity derived from genomic data, one can easily make predictions on phenotypes. For example, the phenotype of a testing sample could be a weighted average of the phenotypes of training samples, where the weights are the similarities between this testing sample and all the training samples. Wheeler et al. (2014) show that their method could provide good predictions of several phenotypes after combining multiple types of genomic data.

**3.4.2. Gene expression regulation modules.** Several integrative genomics methods have been developed to identify gene expression regulation modules. Sun et al. (2007) develop a method to detect modules where a local eQTL modifies the expression of a gene, which modifies the activity of a transcription factor (TF), and the TF in turn regulates the expression of a group of genes. They integrate TF binding-site data and gene expression data to infer latent TF activities and then use genotype data, gene expression, and estimated TF activity to build regulation modules. Akavia et al. (2010) propose a method called CONEXIC (copy number and expression in cancer) to detect modules in which copy number affects the expression of a driver gene, which in turn regulates the expression of a group of genes.

**3.4.3. Study of functional consequences of somatic mutations by integrating somatic mutation data and gene-gene interaction annotations.** Because somatic point mutations tend to be rare, it is difficult to assess their effects directly. Several methods have been developed to borrow information of known gene-gene interactions (e.g., protein-protein interactions or regulation relations) to study the functional consequences of somatic point mutations. The DriverNet method (Bashashati et al. 2012) studies the consequences of somatic mutations on gene expression by connecting genes A and B such that A has a somatic mutation, B has extreme gene expression, and A and B are connected by known gene-gene interaction(s).

Driver mutations that increase the survival advantages of tumor cells often occur together with numerous passenger mutations in cancer patients. Many methods have been developed to find such drivers by identifying recurrently mutated genes. Several recent works have shown that exploiting a mutually exclusive pattern of somatic mutations may help to identify a set of driver mutations. The MEMo (mutual exclusivity modules) method of Ciriello et al. (2012) selects a group of recurrently mutated genes that are close in the gene-gene interaction graph and have mutually exclusive mutations. The focus on genes that are close in the gene-gene interaction graph is partly due to the high computational cost of exhaustive search. However, Leiserson et al. (2013) present a computationally efficient approach to select multiple groups of genes such that the genes within groups have mutually exclusive mutations and good coverage (i.e., most patients have mutations in at least one gene) without relying on gene-gene interaction information.

HotNet (Vandin et al. 2011) identifies significantly altered subnetworks in an interaction network by a network diffusion approach, which can be understood as a random walk on a gene-gene interaction graph. In other words, a somatic mutation in gene A may also affect gene A's neighbors in the interaction graph. After network diffusion, HotNet evaluates the frequency of a subnetwork being altered across patients and find those subnetworks that were recurrently altered. Network diffusion provides a network-smoothed version of the consequences of somatic mutations. Hofree et al. (2013) propose clustering the network-smoothed mutation profiles by nonnegative matrix factorization. The TieDIE method (Paull et al. 2013) uses network diffusion to identify pathways linking somatic mutations and transcriptional regulation pathways.

## SUMMARY POINTS

1. All models have their explicit and implicit assumptions. Model fitting to the underlying data structure largely determines the success of omics data analysis.
2. In data integration, certain biological mechanisms and prior knowledge are often known across different omics data. Proper modeling of such prior knowledge is crucial to enhance statistical power and identify biologically interpretable results.
3. Incorporating prior biological information using Bayesian hierarchical modeling is a very powerful method of data integration.
4. Integration of multiple types of omics data involves ultrahigh-dimensional problems. Feature selection and its related model selection problem are major issues when developing a novel method.
5. Network-based methods are effective approaches to integrating multiple types of data and biological knowledge.

### FUTURE ISSUES

1. Intratumor heterogeneity. Cancer is a somatic evolutionary process and one outcome of such evolutionary processes is that multiple subclones with distinct sets of somatic mutations may co-exist in a tumor sample together with normal cells (e.g., fibroblast cells or blood vessel cells; Beerenwinkel et al. 2015). Most existing methods use somatic mutations and DNA copy number to mathematically deconvolute such mixed signals. Recent studies have shown that gene expression (Yoshihara et al. 2013) and DNA methylation (Zheng et al. 2014) might also be informative, and thus the integrative approach may be useful for cancer subclone studies.
2. Computation scalability. Although data are getting cheaper, computation costs have become more prominent in many applications. Permutation tests are increasingly popular owing to their flexibility, but they are also more computationally costly. In addition, many integrative genomic methods lead to nontrivial optimization problems, and techniques from other fields, such as operations research or machine learning, may provide fruitful avenues to be explored. Developing computationally efficient approaches for integrative genomics is very important.
3. Cross-fertilization between the different styles of integrative approaches, as well as the joint consideration of both horizontal and vertical data integration, will become increasingly relevant.
4. Revolutionary techniques to collect proteomic data from hundreds of thousands of proteins with different combinations of posttranslational modifications are likely to emerge in the near future. Integrative genomics methods to integrate such rich proteomic, transcriptomic, and epigenomic data may greatly improve our understanding of biological systems.

## DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

W.S. is supported in part by US NIH grant R01GM105785. G.C.T. is supported in part by US NIH grant R01CA190766.

## LITERATURE CITED

Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, et al. 2010. An integrated approach to uncover drivers of cancer. *Cell* 143:1005–17

Ancelet S, Abellan JJ, Del Rio Vilas VJ, Birch C, Richardson S. 2012. Bayesian shared spatial-component models to combine and borrow strength across sparse disease surveillance sources. *Biometr. J.* 54:385–404

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene ontology: tool for the unification of biology. *Nat. Genet.* 25:25–29

Bashashati A, Haffari G, Ding J, Ha G, Lui K, et al. 2012. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.* 13:R124

Beerenwinkel N, Schwarz RF, Gerstung M, Markowetz F. 2015. Cancer evolution: mathematical models and computational inference. *Syst. Biol.* 64:e1–e25

Begum F, Ghosh D, Tseng GC, Feingold E. 2012. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res.* 40(9):3777–84

Bergersen LC, Glad IK, Lyng H. 2011. Weighted lasso with data integration. *Stat. Appl. Genet. Mol. Biol.* 10:1–29

Bhadra A, Mallick BK. 2013. Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics* 69:447–57

Bhattacharjee S, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS, et al. 2012. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.* 90(5):821–35

Bottolo L, Chadeau-Hyam M, Hastie DI, Langley SR, Petretto E, et al. 2011a. ESS++: a C++ objected-oriented algorithm for Bayesian stochastic search model exploration. *Bioinformatics* 27:587–88

Bottolo L, Chadeau-Hyam M, Hastie DI, Zeller T, Liquet B, et al. 2013. GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS Genet.* 9:e1003657

Bottolo L, Petretto E, Blankenberg S, Cambien F, Cook SA, et al. 2011b. Bayesian detection of expression quantitative trait loci hot spots. *Genetics* 189:1449–59

Bottolo L, Richardson S. 2010. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Anal.* 5:583–618

Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* 296:752–55

Bühlmann P, Kalisch M, Meier L. 2014. High-dimensional statistics with a view toward applications in biology. *Annu. Rev. Stat. Appl.* 1:255–78

Cai X, Bazerque JA, Giannakis GB. 2013. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Comput. Biol.* 9:e1003068

Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, et al. 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* 30:413–21

Chang L, Lin H, Sibille E, Tseng G. 2013. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinform.* 14:368

Chen LS, Emmert-Streib F, Storey JD, et al. 2007. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol.* 8:R219

Ciriello G, Cerami E, Sander C, Schultz N. 2012. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22:398–406

Cressie NA. 1993. *Statistics for Spatial Data*. New York: Wiley

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, et al. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486:346–52

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, et al. 2004. The Ensembl automatic gene annotation system. *Genome Res.* 14:942–50

Ding L, Wendl MC, McMichael JF, Raphael BJ. 2014. Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet.* 15:556–70

Ellis MJ, Gillette M, Carr SA, Paulovich AG, Smith RD, et al. 2013. Connecting genomic alterations to cancer biology with proteomics: the NCI clinical proteomic tumor analysis consortium. *Cancer Discov.* 3:1108–12

ENCODE Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74

Evangelou E, Ioannidis JP. 2013. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* 14:379–89

Fusi N, Stegle O, Lawrence ND. 2012. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput. Biol.* 8:e1002330

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531–37

Greenberg S, Sanoudou D, Haslett J, Kohane I, Kunkel L, et al. 2002. Molecular profiles of inflammatory myopathies. *Neurology* 59:1170–82

Hageman RS, Leduc MS, Korstanje R, Paigen B, Churchill GA. 2011. A Bayesian framework for inference of the genotype–phenotype map for segregating populations. *Genetics* 187:1163–70

Han B, Eskin E. 2012. Interpreting meta-analyses of genome-wide association studies. *PLOS Genet.* 8(3):e1002555

Hans C, Dobra A, West M. 2007. Shotgun stochastic search for "large *p*" regression. *J. Am. Stat. Assoc.* 102:507–16

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22:1760–74

Hofree M, Shen JP, Carter H, Gross A, Ideker T. 2013. Network-based stratification of tumor mutations. *Nat. Methods* 10:1108–15

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. 2006. RankProd: A bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* 22:2825–27

Huang YT, VanderWeele TJ, Lin X. 2014. Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. *Ann. Appl. Stat.* 8:352

Huo Z, Ding Y, Liu S, Oesterreich S, Tseng GC. 2016. Meta-analytic framework for sparse K-means to identify disease subtypes in multiple transcriptomic studies. *J. Am. Stat. Assoc.* In press

Imholte GC, Scott-Boyer MP, Labbe A, Deschepper CF, Gottardo R. 2013. iBMQ: a R/Bioconductor package for integrated Bayesian modeling of eQTL data. *Bioinformatics* 29:2797–98

Ishwaran H, Rao JS. 2005. Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Stat.* 33:730–73

Jiang Y-h, Bressler J, Beaudet AL. 2004. Epigenetics and human disease. *Annu. Rev. Genomics Hum. Genet.* 5:479–510

Kang DD, Sibille E, Kaminski N, Tseng GC. 2012. MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic Acids Res.* 40:e15

Kendziorski C, Chen M, Yuan M, Lan H, Attie A. 2006. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* 62:19–27

Khatri P, Sirota M, Butte AJ. 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8:e1002375

Kim S, Lin C-W, Tseng GC. 2016. MetaKTSP: A meta-analytic top scoring pair method for robust cross-study validation of omics prediction analysis. *Bioinformatics*. doi: 10.1093/bioinformatics/btw115

Kim S, Xing EP. 2012. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *Ann. Appl. Stat.* 6:1095–117

Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. 2012. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* 28:3290–97

Knorr-Held L, Best NG. 2001. A shared component model for detecting joint and selective clustering of two diseases. *J. R. Stat. Soc. Ser. A* 164:73–85

Kratz A, Carninci P. 2014. The devil in the details of RNA-seq. *Nat. Biotechnol.* 32:882–84

Lee JC, Lyons PA, McKinney EF, Sowerby JM, Carr EJ, et al. 2011. Gene expression profiling of CD8+ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. *J. Clin. Investig.* 121:4170

Leiserson MD, Blokh D, Sharan R, Raphael BJ. 2013. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.* 9:e1003054

Li J, Tseng GC. 2011. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann. Appl. Stat.* 5:994–1019

Li Q, Wang S, Huang C-C, Yu M, Shao J. 2014. Meta-analysis based variable selection for gene expression data. *Biometrics* 70:872–80

Li R, Tsaih SW, Shockley K, Stylianou IM, Wergedal J, et al. 2006. Structural model analysis of multiple quantitative traits. *PLoS Genet.* 2:e114

Lock E, Dunson D. 2013. Bayesian consensus clustering. *Bioinformatics* 29:2610–16

Logsdon BA, Mezey J. 2010. Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Comput. Biol.* 6:e1001014

Malumbres M. 2013. miRNAs and cancer: an epigenetics view. *Mol. Aspects Med.* 34:863–74
Marttinen P, Pirinen M, Sarin AP, Gillberg J, Kettunen J, et al. 2014. Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression. *Bioinformatics* 30:2026–34
Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* 470:59–65
Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, et al. 2013. Pattern discovery and cancer gene identification in integrated cancer genomic data. *PNAS* 110:4245–50
Molitor J, Papathomas M, Jerrett M, Richardson S. 2010. Bayesian profile regression with an application to the national survey of children's health. *Biostatistics* 11:484–98
Monni S, Tadesse MG. 2009. A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Anal.* 4:413–36
Muniategui A, Pey J, Planes FJ, Rubio A. 2013. Joint analysis of miRNA and mRNA expression data. *Brief. Bioinform.* 14:263–78
Neto EC, Ferrara CT, Attie AD, Yandell BS. 2008. Inferring causal phenotype networks from segregating populations. *Genetics* 179:1089–100
Neto EC, Keller MP, Attie AD, Yandell BS. 2010. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Ann. Appl. Stat.* 4:320–29
Newton MA, Wang Z. 2015. Multiset statistics for gene set analysis. *Annu. Rev. Stat. Appl.* 2:95–111
Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12:443–51
Pan W, Xie B, Shen X. 2010. Incorporating predictor network in penalized regression with application to microarray data. *Biometrics* 66:474–84
Papathomas M, Molitor J, Hoggart C, Hastie D, Richardson S. 2012. Exploring data from genetic association studies using Bayesian variable selection and the Dirichlet process: application to searching for gene × gene patterns. *Genet. Epidemiol.* 36:663–74
Pasquinelli AE. 2012. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat. Rev. Genet.* 13:271–82
Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, Stuart JM. 2013. Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (TieDIE). *Bioinformatics* 29:2757–64
Peng J, Zhu J, Bergamaschi A, Han W, Noh D, et al. 2008. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* 4:53–77
Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. 2000. Molecular portraits of human breast tumours. *Nature* 406:747–52
Pettit JB, Tomer R, Achim K, Richardson S, Azizi L, Marioni J. 2014. Identifying cell types from spatially referenced single-cell expression datasets. *PLoS Comput. Biol.* 10:e1003824
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155(2):945–59
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38:904–9
Quintana M, Conti D. 2013. Integrative variable selection via Bayesian model uncertainty. *Stat. Med.* 32:4938–53
Ramasamy A, Mondry A, Holmes CC, Altman DG. 2008. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* 5:e184
Rashid N, Sun W, Ibrahim JG. 2014. Some statistical strategies for DAE-seq data analysis: Variable selection and modeling dependencies among observations. *J. Am. Stat. Assoc.* 109:78–94
Savage RS, Ghahramani Z, Griffin JE, Bernard J, Wild DL. 2010. Discovering transcriptional modules by Bayesian data integration. *Bioinformatics* 26:i158–67
Scott-Boyer M, Imholte G, Tayeb A, Labbe A, Deschepper C, Gottardo R. 2011. An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Stat. Appl. Genet. Mol. Biol.* 11(4):1544–6115
Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28:1353–58

Sheehan N, Didelez V, Burton P, Tobin M. 2008. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Med.* 5:e177
Shen K, Tseng GC. 2010. Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics* 26:1316–23
Shen R, Olshen AB, Ladanyi M. 2009. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25:2906–12
Smith GD. 2007. Capitalizing on Mendelian randomization to assess the effects of treatments. *J. R. Soc. Med.* 100:432–35
Song C, Tseng GC. 2014. Hypothesis setting and order statistic for robust genomic meta-analysis. *Ann. Appl. Stat.* 8:777–800
Song L, Crawford GE. 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* 2010:pdb.prot5384
Spirtes P, Glymour C, Scheines R. 2001. *Causation, Prediction and Search*. Cambridge, MA: MIT Press
Stegle O, Parts L, Durbin R, Winn J. 2010. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* 6:e1000770
Stegle O, Parts L, Piipari M, Winn J, Durbin R. 2012. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* 7:500–7
Stingo FC, Chen YA, Tadesse MG, Vannucci M. 2011. Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes. *Ann. Appl. Stat.* 5:1978–2002
Stingo FC, Chen YA, Vannucci M, Barrier M, Mirkes PE. 2010. A Bayesian graphical modeling approach to microRNA regulatory network inference. *Ann. Appl. Stat.* 4:2024–28
Stirzaker C, Taberlay PC, Statham AL, Clark SJ. 2014. Mining cancer methylomes: prospects and challenges. *Trends Genet.* 30:75–84
Sun W. 2012. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* 68:1–11
Sun W, Hu Y. 2013. eQTL mapping using RNA-seq data. *Stat. Biosci.* 5(1):198–219
Sun W, Wright FA, Tang Z, Nordgard SH, Van Loo P, et al. 2009. Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res.* 37:5365–77
Sun W, Yu T, Li KC. 2007. Detection of eQTL modules mediated by activity levels of transcription factors. *Bioinformatics* 23:2290–97
Terfve C, Cokelaer T, Henriques D, MacNamara A, Goncalves E, et al. 2012. CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Syst. Biol.* 6:133
Thompson JR, Attia J, Minelli C. 2011. The meta-analysis of genome-wide association studies. *Brief. Bioinform.* 12:259–69
Tseng GC, Ghosh D, Feingold E. 2012. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* 40:3785–99
Tyekucheva S, Marchionni L, Karchin R, Parmigiani G. 2011. Integrating diverse genomic data using gene sets. *Genome Biol.* 12:1–14
Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, et al. 2010. Allele-specific copy number analysis of tumors. *PNAS* 107:16910–15
Vandin F, Upfal E, Raphael BJ. 2011. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18:507–22
Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, et al. 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26:i237–45
Wang K, Li M, Hadley D, Liu R, Glessner J, et al. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* 17:1665–74
Wang X, Chua HX, Chen P, Ong RTH, Sim X, et al. 2013. Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* 22:2303–11
Wheeler HE, Aquino-Michaels K, Gamazon ER, Trubetskoy VV, Dolan ME, et al. 2014. Poly-omic prediction of complex traits: OmicKriging. *Genet. Epidemiol.* 38:402–15
Xiong Q, Ancona N, Hauser ER, Mukherjee S, Furey TS. 2012. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Res.* 22:386–97

Yoshihara K, Shahmoradgoli M, Martnez E, Vegesna R, Kim H, et al. 2013. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4:2612

Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B* 68:49–67

Yuan Y, Savage RS, Markowetz F. 2011. Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.* 7:e1002227

Zhang W, Zhu J, Schadt EE, Liu JS. 2010. A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Comput. Biol.* 6:e1000642

Zheng X, Zhao Q, Wu HJ, Li W, Wang H, et al. 2014. MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. *Genome Biol.* 15:1–13

Annual Review of Statistics and Its Application

Volume 3, 2016

# Contents

**Errata**

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics