

Chapter 2

Biostatistics, Data Mining and Computational Modeling

Hao He*, Dongdong Lin*, Jigang Zhang, Yuping Wang,
and Hong-Wen Deng

Abstract With the rapid development of high-throughput experimental technologies, bioinformatics and computational modeling has been a rapid evolving science field concerned with the development of various analysis methods and tools for investigating these large biological data efficiently and rigorously. There are many methods and tools available for the analysis of single omics dataset. It is a great challenge that biological systems are being further investigated by integrating multiple heterogeneous and large omics data. Many powerful methods and algorithmic techniques have been developed to answer important biomedical questions through integrative analysis. In this chapter, in order to help the bench biologist analyze omics data, we introduced various methods from classical statistical techniques for single marker association and multivariate analysis to more recent advances from gene network analysis and integrative analysis of multi-omics data.

Keywords Multi-omics • Integrative analysis • Gene network analysis • Disease diagnosis • Classification

*Author contributed equally with all other contributors

H. He • J. Zhang

Center for Bioinformatics and Genomics, Department of Biostatistics and Bioinformatics,
Tulane University, New Orleans, LA 70112, USA

D. Lin • Y. Wang

Center for Bioinformatics and Genomics, Tulane University, New Orleans, LA, USA

Department of Biomedical Engineering, Tulane University, New Orleans, LA, USA

Hong-Wen Deng, Ph.D. (✉)

Center for Bioinformatics and Genomics, Department of Biostatistics and Bioinformatics,
Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA

e-mail: hdeng2@tulane.edu

2.1 Introduction

In the past decade, with the development of high throughput technologies, massive biological data have been generated from multiple levels of biological systems — including DNA sequence data in genomics, RNA expression levels in transcriptomics, DNA methylation and other epigenetic markers in epigenomics, protein expression in proteomics and metabolic profiling in metabolomics. These omics data are high throughput measurements of the abundance and/or structure features of molecules involved in biological metabolism and regulation. Table 2.1 summarizes the main features of various omics data.

Generally, omics data are high-dimensional data, which means that the number of subject n (e.g., tissue or samples) is much smaller than the number of variables p (e.g., number of SNPs in genome wide association, number of genes in an expression profile). In this setting, we are confronted with thousands of hypothesis testing simultaneously. There is a high risk that statistical models may overfit the omics data. In addition, datasets from diverse genomic levels have unique properties. A better understanding of the data characteristics will help to improve statistical modeling. An increasing number of advanced statistical methods have been developed to address these issues in omics data analysis at different levels.

Table 2.1 Main features of omics data

Omics	Biomarker data	Platforms	Features
Genome	Single nucleotide polymorphism (SNP)	Microarray	Categorical data
	Copy number variation (CNV)	DNA sequencing	Distance-driven correlation
	Loss of heterozygosity (LOH)		Extremely stable over time
	Rare variant		
Transcriptome	Gene expression	Microarray	Continuous data
	Alternative splicing	RNA sequencing	Affected by time and exposures
	Long non-coding RNA		Strong measurement noise
	Small RNA		
Proteome	Protein expression	Microarray	Continuous data
		Mass spectrometry	Affected by time and exposures
Epigenome	DNA methylation	Microarray	Continuous data
	Histone modification	Bisulfite sequencing	Affected by time and exposures
	miRNA		
Metabolome	Metabolite profiling	Mass spectrometry	Continuous data
		Nuclear magnetic resonance (NMR) spectroscopy	Affected by time and exposures
			Structured correlation
			Strongly affected by exposures

Instead of analyzing single omics data, it is interesting to integrate multiple levels of omics data to gain comprehensive insights into biology and disease etiology. It is recognized that multi-scale features do not act in isolation but interact in complex networks (within and across individual omics), e.g., genomics information flow DNA- > RNA- > protein- > traits. Therefore, no single type of omics data can provide a thorough understanding of the complex function/regulatory networks that mediate gene expression/function for disease etiology. Integrative analysis of multiple omics data with the same subjects has the following advantages: 1) multiple omics data can provide diverse information that the identified genetic variants may be consistent in the effects across different omics levels. Consistent results will compensate for unreliable findings in single omics data, which can improve the detection power for those variants with modest effects in individual omics data. Complementary results will confirm the findings to get a more comprehensive understanding of genetic mechanisms of diseases; 2) importantly, integrative analysis of multiple omics data will enable the reconstruction of interplay/regulatory relationship among genetic factors at different levels. The analysis of complex regulatory networks will aid in functional annotation of individual genes/regulatory factors, gaining new insights into the molecular mechanisms underlying disease pathogenesis and generating model hypothesis for further specific testing. Taken together, the integrative trans-omics studies can provide a much more comprehensive view of complex disease etiology than can be achieved by examining individual omics data on their own.

In this chapter, we first briefly review statistical methods for biomarker detection in different omics data. Then we will review integrative statistical analysis involving at least two different types of omics data.

2.2 Statistical Methods for Biomarker Detection in Clinical Bioinformatics

Several types of biological data can be used to identify informative biomarker panels, including SNP data, microarray based gene expression and microRNA. Statistical methods especially predictive models based on these biomarkers are becoming increasingly important in clinical, translational and basic biomedical research. We will first provide illustrations of various statistical methods in the analysis of SNP and gene expression data, attempting to offer practical advice on the appropriate methods to use.

2.2.1 Statistical Analysis for Single Omics Data

2.2.1.1 Single Marker Association

Single SNP Association The objective of genetic association analysis is to establish an association between a phenotype/quantitative trait and a genetic marker.

Usually genetic association tests are performed separately for each individual SNP. A variety of statistical methods could be applied according to the data types of the phenotype/quantitative trait. The phenotype in a study can be case-control (binary), quantitative (continuous), or categorical. First we will discuss analysis for case-control, continuous and categorical disease outcomes and then we will present more advanced statistical methods for multivariate analysis.

Here is the basic problem formulation. Let $\{X_1, \dots, X_p\}$ be a set of P SNPs for N individuals. Suppose the data with each SNP having minor allele a and major allele A . We use 0, 1, 2 to represent the homozygous major allele, heterozygous allele and homozygous minor allele, respectively. Therefore we have $X_{pn} \in \{0, 1, 2\}$, ($1 \leq p \leq P, 1 \leq n \leq N$). Let phenotype be $Y = \{y_1, \dots, y_n\}$. Depending on the data type, the values of Y can be binary, continuous or categorical.

For case-control phenotype, it can be represented as a binary variable with 0 representing controls and 1 representing cases. The association between a SNP and case-control status is to test the null hypothesis of no association between the marker with disease status in a contingency table, which links disease status by either three genotypes counts (A/A, A/a and a/a) or allele count (A and a). The test of association is given by Pearson χ^2 test for the independence of the rows and columns in the contingency table (Balding 2006). The choice of degrees of freedom is based on recessive, dominant and additive models of inheritance. The contingency table can allow alternative models by summarizing the counts based on the models of inheritance. For instance, to test for a dominant model, the contingency table is summarized as 2×2 table of genotype counts (A/A vs. A/a and a/a). As to a recessive model, the contingency table is summarized as 2×2 table of genotype counts (a/a vs. A/A and A/a). There are two tests commonly used for testing the additive model of inheritance: the allele test and the trend test, also known as the Cochran-Armitage trend test. Both tests have the same null hypothesis: $P_{\text{case}} = P_{\text{control}}$, where P_{case} and P_{control} denote the frequency of A alleles among diseased and non-diseased in a population, respectively. As the underlying genetic model is unknown in most genetic association studies, the test for additive model is most commonly used. However, there is no generally accepted answer to the question about what kind of test to be used. The analyses could be designed optimally according to the information that what proportion of undiscovered disease-predisposing variants function additively and what proportions are dominant and recessive. Table 2.2 summarizes different contingency table methods based on diverse tests of association. Take genotypic association for instance, Table 2.3 is the contingency table. For a SNP and the phenotype Y , we use O_{ij} to denote the number of individuals whose X_p equals

i and Y equals j . The Pearson χ^2 statistics is calculated as $\sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, where $E_{ij} = \frac{O_{i.} O_{.j}}{N}$, $O_{i.} = \sum_j O_{ij}$ and $O_{.j} = \sum_i O_{ij}$. The degree of freedom is 2.

Logistic regression is a statistical method for predicting binary and categorical outcome. It can be applied to both single-locus and multi-locus association studies with covariates in the model. Let $Y \in \{0, 1\}$ be a binary variable and $X \in \{0, 1, 2\}$ be

Table 2.2 Tests of association using contingency table methods

Test	DF	Contingency table description
Genotypic association	2	2×3 table of N case-control by genotype counts (A/A vs. A/a vs. a/a)
Dominant model	1	2×2 table of N case-control by dominant genotype pattern of inheritance counts (a/a vs. not a/a)
Recessive model	1	2×2 table of N case-control by recessive genotype pattern of inheritance counts (not A/A vs. A/A)
Cochran-Armitage trend test	1	2×3 table of N case-control by genotype counts (A/A vs. A/a vs. a/a)
Allelic association	1	2×2 table of $2N$ case-control by allele counts (A vs. a)

Note: DF degrees of freedom

Table 2.3 Contingency table for genotypic association test of a single SNP X_p and a phenotype Y

Count	Genotype aa ($X_p = 0$)	Genotype Aa ($X_p = 1$)	Genotype aa ($X_p = 2$)	Total
$Y = 0$ (Control)	O_{00}	O_{01}	O_{02}	$O_{0.}$
$Y = 1$ (Case)	O_{10}	O_{11}	O_{12}	$O_{1.}$
Total	$O_{.0}$	$O_{.1}$	$O_{.2}$	N

a SNP. The conditional probability of $Y = 1$ given a SNP is $\theta(X) = P(Y = 1|X)$. The logit function is defined as $\text{logit}(X) = \ln \frac{\theta(X)}{1-\theta(X)}$. The *logit* function can be taken as a linear predictor function: $\text{logit}(X) \sim \beta_0 + \beta_1 X$. The model can be modified to incorporate multiple SNP loci and potential covariates. For example, the following model fits two predictor SNPs (X_1 and X_2) and two covariates (Z_1 and Z_2): $\text{logit}(X) \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z_1 + \beta_4 Z_2$.

For continuous (quantitative) traits, the basic statistical tools are linear regression and analysis of variance (ANOVA).

In **regression models**, there are two types of variables: dependent variable (response variable or outcome variable) and independent variable (explanatory variable or predictor variable). In a regression model, the dependent variable is modeled as a function of one or more independent variables. When this function is a linear combination of one or more model parameters, called regression coefficients, the model is called a linear regression model. A least-squares regression line is often used to find optimal fit between the phenotype and the genotype.

For simplicity, a single SNP genotype is denoted X_i and the phenotype is $Y_i, i = 1, \dots, n$. For this given data set (X_i, Y_i) , we are fitting a simple linear regression model, $Y = \beta_0 + \beta_1 X + \varepsilon$, such that $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2$, and

ε 's are uncorrelated. We can find b_0 and b_1 as least squares estimators for β_0 and β_1 , respectively. We have the sums of squares as follows: $S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2$,

$$S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2, \text{ and } S_{XY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}), \text{ and the following two normal equations, } b_0 + b_1 \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i \text{ and } b_0 \sum_{i=1}^n X_i + b_1 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i.$$

The estimator of b_1 is $\frac{S_{XY}}{S_{XX}}$. Then we can test the null hypothesis against the alternative hypothesis $H_0 : \beta_1 = \beta_{10}$ versus $H_1 : \beta_1 \neq \beta_{10}$, where β_{10} is a specified value that could be zero. The test statistics is calculated as

$$t = \frac{(b_1 - \beta_{10})}{se(b_1)} = \frac{(b_1 - \beta_{10}) \left\{ \sum_{i=1}^n (X_i - \bar{X})^2 \right\}^{\frac{1}{2}}}{\sqrt{S^2}}, \text{ where } S^2 \text{ is the estimate of residual mean}$$

square $\sigma_{Y.X}^2$. One can compare $|t|$ with $t(n-2, 1-\frac{\alpha}{2})$ from a t-table with $(n-2)$ degrees of freedom. The test is a two-sided test conducted at the $100\alpha\%$ level.

In one-way ANOVA the F-test is used to assess whether the expected values of a quantitative variable within several pre-defined groups differ from each other. For a single SNP, we can divide all the subjects into three groups according to their genotypes. Let $Y'_i (i \in \{0, 1, 2\})$ be the subset of phenotypes for the subjects corresponding to genotype i . The number of subjects with Y'_i is denoted as n_i . Note that $\sum_{i=0}^2 n_i = N$. The total sum of squares (SST) can be divided into two parts, the between-group sum of squares (SSB) and the within-group sum of squares (SSW).

$$SSB = \sum_{i=1}^2 (\bar{Y}'_i - \bar{Y})^2, SST = \sum_{i=0}^2 \sum_{n=1}^N (Y'_{in} - \bar{Y})^2, \text{ and } SSW = SST - SSB. \text{ The}$$

formula of F-test statistic is $F = \frac{SSB}{SSW}$, and F follows the F-distribution with 2 and $N-3$ degrees of freedom under the null hypothesis.

Gene Expression Analysis In transcriptomics studies for biomarker discovery among thousands of features, we are interested in which genes/features are differentially expressed under two (or more) conditions. The hypothesis test will be performed individually for each feature. Statistical significance for each hypothesis test is assessed according to its corresponding p-value from a statistical test. Suppose there are K conditions and n_k samples in the k th condition in a total of N samples, where $K \in \{1, 2\}$. Let X_{ijk} be an expression value, where sample $i = 1, 2, \dots, n_k$, gene features $j = 1, 2, \dots, m$, and condition $K = 1, 2$. Assume that gene expression values have been background corrected, normalized and transformed by taking the logarithm to base 2. The sample mean and variance of

gene feature j in group k are given as $\bar{X}_{jk} = \frac{\sum_{i=1}^{n_k} X_{ijk}}{n_k}$ and $S_{jk}^2 = \frac{\sum_{i=1}^{n_k} (X_{ijk} - \bar{X}_{jk})^2}{n_k - 1}$, respectively.

Fold change approach is a simple and straightforward way of evaluating the degree of differential expression under two conditions. For a gene feature j , the mean difference is given by $M_j = \bar{X}_{j1} - \bar{X}_{j2}$. Then the fold change is a statistic 2^{M_j} . Gene will be declared as significant if $|M_j|$ is greater than a predefined threshold. Such procedure assumes that the variances are equal across all genes. However, it is not the case for gene expression profile. Therefore, this approach may easily yield many false positive and false negative results in differential expression analyses.

The two-sample t-test is a most used parametric statistical test in differential expression analysis. It compares the means of expression value in two groups taking the variance into consideration. Statistically, we want to test the null hypothesis $H_0 : \mu_{j1} = \mu_{j2}$ against the alternative hypothesis $H_1 : \mu_{j1} \neq \mu_{j2}$ for $j = 1, 2, \dots, m$. The test

statistic for each j is $t_j = \frac{\sum_{i=1}^n (\bar{X}_{j1} - \bar{X}_{j2})^2}{s_j}$, where $s_j = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1-1)S_{j1}^2 + (n_2-1)S_{j2}^2}{n_1+n_2-2}}$,

called pooled within-group standard error. Under the null hypothesis, t_j follows Student's t -distribution with $n_1 + n_2 - 2$ degrees of freedom. A p -value can be found using a t -distribution table. By using the pooled within-group standard error estimated from each gene separately, the t -statistic takes into consideration of variance across different genes.

Significance analysis of microarrays (SAM) is a statistical technique for determining whether changes in gene expression are statistically significant (Tusher et al. 2001). In SAM, statistically significant genes will be identified based on gene specific t -tests. A statistic d_j for each gene j measures the strength of the relationship between gene expression and a response variable. Non-parametric statistics is used as the data may not follow a normal distribution. SAM will perform repeated permutations for the data to determine the significance of any gene with the response. The use of permutation-based analysis accounts for correlations in genes and avoids parametric assumptions about the distribution of individual genes. It assumes equal variance and/or independence of genes. This is an advantage over other techniques. Here is the generic procedure for SAM. A statistic d_j is computed as $d_j = \frac{r_j}{s_j + s_0}$, where r_j is a score, s_j is a standard deviation and s_0 is an exchangeability factor. Compared with the standard t -statistic, the SAM's procedure adds a s_0 term to the denominator. The rationale behind it is that the variance s_j tends to be smaller at lower expression levels, making d_j dependent on the expression levels. However, in order to compare d_j across all genes, the distribution of d_j should be independent of the expression levels. Therefore, SAM seeks to find a s_0 such that the dependence of d_j on s_j is as small as possible. An appropriate value of s_0 will be picked such that the coefficient of variation of d_j is approximately constant as a function of s_j . For details of the SAM procedure, please refer to the tutorial document for the software package, SAM, at <http://statweb.stanford.edu/~tibs/SAM/sam.pdf>.

The **Wilcoxon rank-sum test**, also known as the Mann–Whitney U-test, is a nonparametric test, which can be applied to data with unknown distributions contrary to t -test applied only to normal distributions. It is nearly as efficient as the t -test on normal distributions. The null hypothesis of the test is that two samples come from the same population and an alternative hypothesis is that a particular population tends to have larger values than the other. The Wilcoxon rank-sum test is based on the ranks of the original data values. To perform the Wilcoxon rank-sum test, one first assigns numeric ranks to all the observations, beginning with 1 for the smallest value. Where there are groups of tied values, assigning a rank equal to the midpoint of unadjusted rankings. Second, one adds up the ranks for the observations which came from group 1. The sum of ranks in group 2 is now determinative, since the sum of all the ranks equals $N(N+1)/2$ where N is the total number of observations. Then calculate $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ and $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$. The smaller value of U_1 and U_2 is the one used when consulting significance tables.

2.2.1.2 Multiple Testing

As mentioned earlier, in omics studies we are confronted with a great number of hypotheses to be tested simultaneously. It will result in an inflation of the family wise error rate (FWER) if there is no adjustment for multiple tests. In statistical hypothesis testing, a type I error occurs when the null hypothesis (H_0) is true, but is rejected (a “false positive”). A type II error occurs when the null hypothesis is false, but erroneously fails to be rejected (a “false negative”). A type I error is the incorrect rejection of a true null hypothesis (a “false positive”), while a type II error is the failure to reject a false null hypothesis (a “false negative”). Basically, in hypothesis testing, we want to maximize the power (=1-the type II error) while controlling the type I error less than or equal to a predetermined significance level α . In particular, consider the problem of testing simultaneously m null hypothesis H_j : no differential expression against H_j^a : differential expression, where $j = 1, 2, \dots, m$. A gene will be considered as significantly differentially expressed if its p-value is less than the defined significant level α . However, for hypothesis testing, the problem of multiple testing problem results from the increase in type I error that occurs when many statistical tests are used simultaneously. Suppose there are m independent comparisons, the experiment-wide significance level $\bar{\alpha}$, also termed FWER, is given by $\bar{\alpha} = 1 - (1 - \alpha)^m$. $\bar{\alpha}$ increases as the number of comparison increases. Multiple testing correction is to re-calculate the probabilities obtained from a statistical test which was repeated multiple times. In order to retain FWER $\bar{\alpha}$ in an analysis, the error rate for each comparison must be more stringent than α .

A number of procedures for controlling error rates have been developed to solve the multiple-testing problem. One of the most commonly used approaches for multiple comparisons is the Bonferroni procedure for controlling the FWER at level α , which rejects any hypothesis H_j with unadjusted p-value less than or equal to α/m . The Bonferroni procedure is very conservative. A less conservative

procedure is the Benjamini–Hochberg procedure (BH step-up procedure), which controls the false discovery rate (at level α). The procedure works as follows: first for a given α , find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha$. Second, reject all H_j for $j = 1, 2, \dots, k$. The BH procedure is valid when the m tests are independent and also in various scenarios of dependence.

2.2.1.3 Multivariate Analysis

Although many common genetic variants associated with complex traits have been identified by GWAS, these traits are typically analyzed separately in a univariate manner for association with DNA markers. However, multivariate analysis for correlated traits could be very advantageous in several aspects. First, when there is genetic correlation between different traits, a multivariate analysis can increase power by using the extra information provided by the cross-trait covariance, which is ignored by the univariate analysis. Second, a multivariate analysis of multiple traits can reduce the number of performed tests and alleviate multiple testing burden compared to analyzing all traits separately. Lastly, a multivariate analysis is biologically making more sense as a single genetic marker is associated with multiple traits, compared to the cross-trait comparison in univariate analysis (Galesloot et al. 2014).

A number of multivariate analysis methods in population-based GWAS have been published. Here we briefly introduce six methods including as well as their softwares.

The multivariate test of association MQFAM is implemented in the genetic association analysis software PLINK (MV-PLINK) (Ferreira and Purcell 2009; Purcell et al. 2007). The command used for association testing with MV-PLINK (<https://genepi.qimr.edu.au/staff/manuelF/multivariate/main.html>) is: `plink.multivariate -noweb -file geno -mqfam -mult-pheno pheno.phen -out output`. For each genetic variant, MV-PLINK produces an F-statistic and a p-value in the additive model. Canonical correlation analysis (CCA), which is a multivariate generalization of the Pearson product-moment correlation, to measure the association between the two sets of variables. Specifically, CCA extracts the linear combination of traits that explain the largest possible amount of the covariation between the marker and all traits. The interpretation of a significant multivariate test is aided by the inspection of the weights attributed by the CCA to each phenotype.

Bayesian multiple phenotype test is implemented in SNPTEST (MV-SNPTEST) (Marchini et al. 2007). The command used to perform additive association testing with MV-SNPTEST is provided in the online tutorial (https://mathgen.stats.ox.ac.uk/genetics_software/snpctest/snpctest.html#multiple_phenotype_tests). The model is the Bayesian Multivariate Linear model which is specified by $(y_{i1}, \dots, y_{iq})^T = G_i(\beta_1, \dots, \beta_q)^T + (e_{i1}, \dots, e_{iq})^T$, where $(e_{i1}, \dots, e_{iq})^T \sim N(0, \Sigma)$ and (y_{i1}, \dots, y_{iq}) is the vector of the q residual phenotypes measured on the i th

individual. G_i is the code of the SNP genotype for the i th individual. We use the conjugate prior for this model. This is an inverse Wishart prior $IW(c, Q)$ on the error covariance matrix Σ and a matrix normal (N) prior on the vector of parameters $(\beta_1, \dots, \beta_q) \sim M \sim N(V, \Sigma)$, where M is a mean vector and V is a constant. An inverse Wishart prior $[IW(6, 4)]$ was set on the error covariance matrix Σ and a matrix normal prior $[N(0.02, \Sigma)]$ on the vector of parameters, according to recommendations of the authors. Method ‘expected’ will result in the use of expected genotype counts (\sim dosages) in the analyses.

MultiPhen is an R package available from CRAN (<https://cran.r-project.org/web/packages/MultiPhen/index.html>) (O’Reilly et al. 2012). The regression performed at a SNP, g , and a phenotype, k , to test for association between the SNP genotypes and the phenotype is: $Y_{ik} = \alpha_k + \beta_{gk}X_{ig} + \varepsilon_{igk}$, where ε_{igk} is the residual error assumed to be normally distributed. The null hypothesis of no association between SNP and genotype can be tested by performing a t-test on the null hypothesis $\beta_{gk} = 0$. In the MultiPhen approach, the regression is inverted so that the SNP genotype, X , becomes the dependent variable, and K phenotypes under study become the predictor variables. The genotype data is an allele count and is therefore modelled using ordinal regression; we use proportional odds logistic regression. This model defines the class probabilities as follows.

$$P(X_{ig} \leq m) = \frac{1}{(-\alpha_{gm} - \sum_{k=1}^K \beta_{gk} Y_{ik})}$$

tion is a likelihood ratio test (LRT) for model fit, testing the null hypothesis $\beta_{g1} = \dots = \beta_{gk} = 0$. This results in a p value per trait and a p-value for the LRT.

A Bayesian model comparison and model averaging for multivariate regression is implemented in BIMBAM software (Stephens 2013). The details of statistical method are provided in the reference (Stephens 2013). The BIMBAM software can be run in two different ways. First we test for association between the multivariate traits, all partitioned in the group of directly affected traits, and genotype. Second, we consider all possible partitions of traits into the different categories of traits (directly affected, indirectly affected, and unaffected).

The Principal Component of Heritability Association Test (PCHAT) (Klei et al. 2008) is implemented in the software available at <http://www.wpic.pitt.edu/wpiccompngen/PCHAT/PCHAT.htm>. First, the sample is split into a training set and a test set. The training set is used to construct the optimal linear combination of traits from a heritability point of view. A test set is used for association testing between genotype and the optimal linear combination of traits. In this way, use of the same data for both estimation of the optimal linear combination of traits and association testing is avoided. In addition, a ‘bagging’ approach is performed, in which bootstrap samples are drawn from the training sample and the optimal linear combination of traits is averaged across bootstrap samples. The null distribution of the test statistic is obtained in the same way, using permutation of the data.

A Trait-based Association Test (TATES) is based on Extended Simes procedure (van der Sluis et al. 2013). TATES (<http://ctglab.nl/software>) constitutes a powerful new multivariate strategy that allows researchers to identify novel causal variants. TATES acquire one trait-based p-value by combining p-values in standard univariate GWAS, while correcting for correlations between components. It can detect both genetic variants which are common to multiple phenotypes and those which are specific to a single phenotype. It requires a correlation matrix of the traits and univariate association results as input. The *corr* function in R can be used to generate the full and symmetrical correlation matrices. TATES was run in R and the output contains the TATES trait-based p-value corrected for the correlations between the traits.

2.2.1.4 Gene Set Analysis

In transcriptomics study, massive throughput techniques, such as microarray and RNA sequencing, allow to identify differentially expressed genes (DEGs) associated with diseases or phenotypes from genome-wide gene expression profile. The challenge in expression data analysis in recent years has shifted from single DEG analysis to gene set analysis (GSA), as biologically many complex diseases may be modestly regulated by a set of related genes rather than a single gene. The gene sets are defined based on prior biological knowledge, e.g., biochemical pathways or coexpression in previous experiments. GSA can alleviate the difficulty in interpretation of multiple testing lists of DEGs and provide insights into biological mechanisms for complex diseases. The first and most popular GSA is gene set enrichment analysis (GSEA) (Subramanian et al. 2005), which is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes). The GSEA method is implemented in a freely available software package at <http://www.broadinstitute.org/gsea/index.jsp>. The basic idea for this method is presented as follow (Subramanian et al. 2005):

- Step 1: Calculate an Enrichment Score. Rank genes by their expression difference in two biological states and then compute cumulative sum over ranked genes. The magnitude of increment depends on correlation of gene with phenotype. Record the maximum deviation from zero as the enrichment score.
- Step 2: Estimate significance. Permute phenotype labels 1000 times and compute ES score for each permutation. Then compare ES score for actual data to distribution of ES scores from permuted data.
- Step 3: Multiple Hypothesis Testing. Normalize the ES accounting for size of each gene set to obtain the normalized enrichment score (NES). Calculate FDR for each NES to control proportion of false positives by comparing tails of the observed and null distributions for the NES.

Another interesting GSA method proposed by Efron and Tibshirani attempts to combine gene and sample randomization in one procedure (Efron and Tibshirani

2007). It shows that it is more powerful based on the “maxmean” statistic than the modified Kolmogorov-Smirnov statistic used in GSEA. This method can be implemented by the R package “GSA”. The basic procedures are summarized here:

1. Compute a summary statistic z_i for each gene, for example the two sample t-statistic for two-class data. Let z_s be the vector of z_i values for genes in a gene-set S .
2. For each gene-set S , choose a summary statistic $S = s(z)$: the maxmean statistic
$$\left\{ \left| \frac{\sum_{i=1}^m I(z_i > 0) z_i}{m} \right|, \left| \frac{\sum_{i=1}^m I(z_i < 0) z_i}{m} \right| \right\}$$
3. Standardize S by its randomization mean and standard deviation as $S' = \frac{(S - \text{mean}(s))}{\text{std}(s)}$. For summary statistics such as the mean, mean absolute value or maxmean, this can be computed from the genewise means and standard deviations, without having to draw random sets of genes.
4. Compute permutations of the outcome values (e.g., the class labels in the two-class case) and re-compute S' on each permuted dataset, yielding permutation values. Use these permutation values to estimate p-values for each gene-set score S' and false discovery rates applied to these p-values for the collection of gene-set scores.

In 2007, Wang et al. extended the GSEA to GWAS of complex diseases (Wang et al. 2007), where multiple genes in the same GS/pathway contribute to disease etiology but where common variations in each of those genes make modest contributions to disease risk. Gene set analysis tests disease association with genetic variants in a group of functionally related genes, such as those belonging to the same biological pathway. It can potentially improve the power to detect causal GS/pathways and disease mechanisms by considering multiple contribution factors together, rather than focusing on the top SNPs associated with disease. Individual SNPs in univariate analysis only account for a small proportion of the heritability of complex diseases. The method assesses the enrichment of significant associations for genes in the GS/pathway (as compared with those outside the GS/pathway) using a weighted Kolmogorov–Smirnov running-sum statistic. The GSEA method is modified to fit GWAS data. For each SNP V_i ($i = 1, \dots, L$, where L is the total number of SNPs in a GWA study), its test statistic value is calculated, r_i (e.g., a χ^2 statistic for a case-control association test). We next associated SNP V_i with gene G_j ($j = 1, \dots, N$, where N is the total number of genes represented by all SNPs) if the SNP is located within or <500 kb away from the gene. The highest statistic value among all SNPs mapped to the gene, is assigned as the statistic value of the gene. For all N genes that are represented by SNPs in the GWA study, their statistic values are sorted from largest to smallest, denoted by $r_{(1)}, \dots, r_{(N)}$. For any given gene set S , composed of N_H genes, a weighted Kolmogorov-Smirnov-like running-sum statistic is calculated which reflects the overrepresentation of genes within the set S at the top of the entire ranked list of genes in the genome.

Over recent years, various methods have been published for gene-set or pathway-based association analysis for GWAS. Basically, these statistical methods can be classified into two categories based on whether the required input data sets are a collection of SNP p -values or individual-level SNP genotypes. Additionally, the null hypothesis can also be categorized as ‘self-contained’ versus ‘competitive’ based on whether comparisons were made between genes in a specific pathway and non-associated genes or other genes in the genome. Some of these published algorithms as well as software implementations or web servers are summarized in the review (Wang et al. 2010).

2.2.1.5 Gene Network Analysis

Recent years many network theories have been applied to gene coexpression network analysis. As gene expression microarrays measure the transcription levels of thousands of genes simultaneously, it provides great opportunities to explore large scale gene regulatory networks. Genes with similar expression patterns may participate in pathways and in regulatory and signaling circuits and their products may form complexes. Gene networks provide a systematic understanding of molecular mechanisms underlying biological processes, and the visualization of direct dependencies facilitates systematic interpretation and comprehension of the relationships among genes. Most complex human diseases are arising not from a single gene but from interactions with many other genes, especially in a gene network. The hub genes, which interact with many other genes, are likely to be drivers of the disease status. The analysis on the hub genes has become a promising approach for identifying the key candidate genes for complex diseases.

A great number of statistical methods for gene network reconstruction from gene expression microarray data have been proposed in recent years. There are four main categories of statistical methods: (1) Probabilistic networks-based approaches, mainly Bayesian networks (BN), (2) correlation-based methods, (3) partial-correlation-based methods, and (4) information-theory-based methods (Allen et al. 2012). The representative method in each category and the implementation software are summarized below.

Probabilistic networks, mainly Bayesian networks, are based on a probabilistic graphical model that represents a set of variables and their probabilistic independencies. The Bayesian networks expand the joint probability in terms of simpler conditional probabilities, which allow them to handle noise inherent in both biological processes and microarray experiments. Generally, the joint likelihood function of nodes X_1, \dots, X_p in a Bayesian network can be expressed as

$$P(X_1, \dots, X_p) = \prod_{i=1}^p P(X_i | \prod_i^G),$$
 where graph $G = (V, E)$ represents the topological structure of the Bayesian network, in which $V = \{X_1, \dots, X_p\}$ denotes the set of nodes and $E = \{X_j \rightarrow X_i, X_j \in \prod_i^G\}$ denotes the set of edges. Werhli’s

implementation for Bayesian network construction method is most used and outperforms other implementations (Werhli et al. 2006). A Bayesian network models the distribution of observations and a causal network models the distributions of observations and effects of interventions. A causal network can be interpreted as a Bayesian network, when we are willing to make the Causal Markov Assumptions: given the values of a variable's immediate causes, it is independent of its earlier causes (Friedman et al. 2000).

Correlation-based methods are the most straightforward and popular way to explore the gene co-expression network. They have been successfully applied in many studies and have shown their usefulness in identifying important gene modules and in interpreting biological results. Basically a gene co-expression similarity matrix is defined as $S = [S_{ij}]$, where S_{ij} is the pair-wise transcription correlation coefficients between gene i and j . S is the correlation matrix (Zhang and Horvath 2005). Particularly, Weighted Correlation Network Analysis (WGCNA) is a representative method for the correlation-based approach (Langfelder and Horvath 2008). The implementation of WGCNA is in R package, which is used for identifying modules/subnetworks using hierarchical clustering approaches. The WGCNA R package includes interfaces with Cytoscape (Shannon et al. 2003) for network visualization and The database for annotation, visualization and integrated discovery (DAVID) (Dennis et al. 2003) for enrichment analysis. The comprehensive set of online tutorials that guide users through the major steps for gene network analysis by WGCNA are provided in the website <http://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/index.html>. In the tutorials, R code in each step is provided so that the user can copy and paste into an R session. The tutorials cover the following major topics: correlation network construction, step-by-step and automatic module identification, consensus module detection, eigengene network analysis and differential network analysis.

Here we briefly review the key concepts of the WGCNA framework. The nodes in a gene coexpression network correspond to genes, labeled by indices $i, j = 1, 2, \dots, n$. The edge between two nodes is determined by the pairwise correlation. The network can be specified by its *adjacency matrix* \mathbf{A} , a symmetric matrix with entries a_{ij} in $[0, 1]$ that encode the strength of the link between genes i and j . An unsigned network is defined by the adjacency \mathbf{A} in terms of *coexpression similarity* $S_{ij} = |cor(x_i, x_j)|$, in which positive and negative correlations are treated equally. Also if we want to preserve the sign of the correlation, we can use a *signed* similarity defined as $S_{ij} = \frac{(1 + cor(x_i, x_j))}{2}$. The main difference between signed and unsigned similarities is that genes with a high negative correlation (close to -1) will have a low similarity in a signed network but a high similarity in an unsigned network. A weighted network can preserve the continuous nature of the co-expression information by using a soft thresholding parameter, $\beta \geq 1$. By using a power function, the connection strength can be assessed, $a_{ij} = S_{ij}^\beta$. The default values $\beta = 6$ and $\beta = 12$ are used for unsigned and signed networks.

In WGCNA, genes are clustered into network modules based on their coexpression. Highly coexpressed genes have a small dissimilarity. For example, the adjacency-based dissimilarity measure is $dissAdj_{ij} = 1 - a_{ij}$. The dissimilarity measure can be used as input in average linkage hierarchical clustering. Then, modules are defined as branches of the resulting cluster tree. If larger and more robust modules are desired, one can use a dissimilarity measure based on the topological overlap matrix (TOM):

$$dissTOM_{ij} = 1 - TOM_{ij} = 1 - \frac{\sum_{u \neq i} a_{iu}a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}, \text{ where } k_i = \sum_{u \neq i} a_{ui} \text{ denotes the}$$

network connectivity. TOM combines the connection strength between a pair of genes with their connections to other ‘third party’ genes, which has been shown to be a highly robust measure of network interconnectedness (proximity). In order to summarize the module genes by a single representative expression profile, module eigengene is defined as the first principal component of the standardized expression profiles of a given module, which is considered as the weighted average of the module gene expressions. We can correlate the module eigengenes with the trait of interest y . The correlation coefficient or its corresponding p-value is referred to as the eigengene significance. For each module, the module significance is defined as the average absolute gene significance for all genes in the module. WGCNA can alleviate the multiple testing problem in DEG analysis, as it focuses on a few modules with the trait rather than thousands of genes and these modules may be included into some important biological pathways.

Partial-correlation-based methods are based on Gaussian graphic model. These methods infer the conditional dependency by the non-zero entries in the precision matrix, $C = [C_{i,j}] = S^{-1}$, which is the inverse of covariance matrix (Allen et al. 2012). The zero entries in the precision matrix imply conditional independency between the expression levels of gene i and j given the expression of all other genes, which means two genes do not interact directly with each other. The sparse partial correlation estimation (SPACE) algorithm is a representative partial-correlation-based method (Peng et al. 2009). It converts the concentration matrix estimation problem to a regression problem and optimizes the results with a symmetric constraint and an L_1 penalization.

Information-theory-based methods use mutual information (MI) to determine how similar the joint distribution $P(X, Y)$ is to the products of factored marginal distribution $P(X)P(Y)$. It can determine the dependency among the genes and then remove indirect interactions. Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) is a successful and popular information-theory-based method, which has been successfully applied to construct gene regulatory networks in the context of specific cellular types (Margolin et al. 2006). The calculation of MI does not assume a monotonic relationship; therefore it is able to identify the non-linear or irregular dependencies, which will be missed by Pearson correlation. If the gene network contains non-monotonic dependencies the ARACNE could outperform correlation-based methods.

2.2.2 Computational Methods for Integrating Multi-Omics Data

A variety of statistical methods and tools have been proposed for integrating two or more omics data. These methods aim to help understand molecular mechanism or biological pathways underlying variation of different types of clinical traits. Also they explore the relationship or interactions among diverse omics data for complex network structure reconstruction and thereby identifying risk modules associated with clinical outcomes. Integrated information is finally used for subtyping clinical diseases or predicting the outcome for prospective patients. These computational methods can be broadly categorized into four types in terms of the objective of analysis and the way of integrating omics data.

2.2.2.1 Multi-Stage Method: Analyzing Multi-Omics Data Sequentially

Multi-stage method is a way to divide multi-omics analysis into multiple stages, where each stage only incorporates two levels of omics and subsequently relates biomarkers to the trait or phenotype of interest. For example, a three stage strategy is commonly applied for identifying genetic variants associated with the phenotype and relating the other levels of omics, e.g., gene expression (Holzinger and Ritchie 2012).

- Step1. Identifying those significant genetic variants (e.g., SNPs) associated with phenotype by genome-wide association test with multiple testing corrected.
- Step2. Testing those identified SNPs for association with the other omics data, such as gene expression, DNA methylation, protein expression and other functional profiling. The corresponding associated SNPs are called expression quantitative loci (eQTLs (Jansen and Nap 2001)), methylation QTL (meQTLs (Kerkel et al. 2008)), protein QTL(pQTLs (Melzer et al. 2008)) respectively.
- Step3. Those omics features having at least one QTL are further tested for the association with phenotype. Subsequently, biological pathways can be derived; some SNPs associate with phenotype through other omics data while some SNPs can affect phenotype independent of the other omics data. One benefit of multi-stage method is that each single stage analysis is performed independently with a variety of statistical methods (Cantor et al. 2010). For example, to identify significant biomarkers at the first and third stage, both univariate test (e.g., linear regression or logistic regression) and multivariate methods (e.g., region or pathway based test (Khatri et al. 2012)) can be applied for genome-wide detection. At the second stage, many approaches proposed for identifying eQTLs can also be applied for the analysis of meQTLs, or pQTLs, such as single-trait QTL tests, multi-trait QTL methods, and QTL test with pedigree or error correction (Kendzioriski et al. 2006).

Some multi-stage methods have been proposed for sequential analysis of multi-omics data. For instance, Schadt et al. applies multistep method to analyze DNA methylation, gene expression and other complex traits to determine if the variation of DNA methylation that leads to the change of gene expression traits statistically supports an independent, causative or reactive function relative to the complex traits (Schadt et al. 2005). Hao et al. performed a systematic analysis and identified two modules underlying BMD by incorporating GWASs, human PPI network, and gene expression (He et al. 2014). The tool, Multiple Concerted Disruption (MCD) is proposed to sequentially search for a set of genes which exhibit concerted disruption through multiple genomic dimension (DNA methylation, copy number and allelic status) and consequential change in gene expression (Chari et al. 2010). The procedure involves four sequential steps with increasing number of genomic data incorporated to filter out those genes lacking concerted disruption. Similar method for exploring the relationship between copy number alternation and methylation (CNAmet) is also proposed (Louhimo and Hautaniemi 2011). In addition, prior knowledge such as KEGG pathway, gene ontology or functional annotation of the region (e.g., transcription factor binding, methylated or regulatory motifs) could also be incorporated into the analysis to refine the specific regions of interest for the subsequent multi-stage analysis.

Although it is easy to model the relationship among multi-omics data by exploring their pair-wise relationship sequentially, there is a limitation for the stepwise hypothesis. If different omics interplay to have joint effect, for example, miRNA and DNA methylation may simultaneously affect the gene expression, the multi-stage methods may lose their efficiency.

2.2.2.2 Parallel Analysis: Combining Individual Omics Analysis Results

Parallel analysis combines multi-omics data into the analysis simultaneously. It can be generally divided into two categories: concatenation-based integration and model-based integration.

Concatenation-Based Integration This method is to straightforwardly concatenate all of omics data from the same subjects, resulting in a large combined matrix. One advantage of this integration is the applicability of many single omics analysis methods if combining features appropriately. For example, a variety of univariate and multivariate association tests could be applied for biomarker detection from different levels of features, especially the penalized likelihood methods which can handle high dimensionality of data. Lasso is a very useful penalized method and has been widely used for feature selection (Tibshirani 1996). Recently significant test based on lasso is also proposed to control the type I error (Lockhart et al. 2014). Other penalized methods such as sparse logistic regression (Shevade and Keerthi 2003), cox lasso (Wang et al. 2009), and sparse multinomial regression (Krishnapuram et al. 2005) have also been used for genetic biomarker identification corresponding

to different types of phenotypes (e.g., categorical or survival traits). These methods can be extended to the analysis of concatenated matrix consisting multi-omics data.

Another advantage of concatenating datasets is that they can account for relationship among features from different levels of omics data. For example, SNP and DNA methylation measure the effect of genetic mutation and environmental factors on complex traits respectively. They may interact with each other to deregulate gene expression, leading to the variation of traits. Fridley et al. used Bayesian modeling to incorporate the relationship between SNPs and mRNA gene expression into the concatenation-based association model for the prediction of drug cytotoxicity (Fridley et al. 2012). In penalized likelihood methods, elastic net is used to simultaneously select features and account for the correlation among features (Ogutu et al. 2012). Group based penalties (e.g., group lasso, sparse group lasso, group Bridge, and overlapping group lasso) were proposed to group different levels of features based on their genomic annotation (e.g., gene or pathway) to increase the detection power on group level (Huang et al. 2012). In addition, Lando et al. used the correlation between copy number and phenotype to weight the penalty of gene expression in a penalized regression model. Genes corresponding to important CNVs were less penalized in expression regression model (Lando et al. 2009).

In spite of the advantages of concatenating multi-omics data, it is still a challenge to find an appropriate way to combine these data matrices collected from different platforms with different scales into one model. In addition, the combination of these high-dimensional matrices will largely expand the dimension of the model, which could increase computational burden. Therefore, the concatenation of multiple datasets is more applicable for omics data integration if there exists an appropriate way of concatenating matrix and the dimension of data is moderate.

Model-Based Integration To avoid the issues of combining data directly, some studies try to build a model for each data separately and then transform each model into an intermediate form, and finally integrate transformed outputs for multi-omics analysis. Tyekucheva et al. performed gene-level and gene set-level tests on gene expression and copy number data separately and combined the gene set scores by meta-analytical approaches (e.g., geometrically averaged P-values and minimum P-values) to derive the combined gene-set score (Tyekucheva et al. 2011). The integrative approach identified more reliable glioblastoma multiforme tumor related gene sets than individual data analysis. Similarly, Poisson et al. proposed the sum of square statistics to combine gene set score from gene expression and metabolites to test integrative set enrichment (Soneson et al. 2010). Xiong et al. developed a tool, Gene Set Association Analysis (GSAA), to test gene-set enrichment by combining SNP-set and gene expression using different score based combination methods (e.g., z-score sum, rank sum and fisher's test) (Xiong et al. 2012). Analysis Tool for Heritable and Environmental network Association (ATHENA) is another model-based analysis tool for performing integrative analysis of different omics data as well as their association with clinical outcomes (Holzinger et al. 2013).

Besides the statistical model or score integration, multi-task learning is another powerful strategy to jointly model different but related tasks simultaneously. Biomarker identification in each single omics is treated as a task and then multiple tasks are combined by multitask learning. Bennett et al. used multi-task learning to consider enrichment analysis scores from both SNP and gene expression to identify several pathways with both genetic and expression differences related to the phenotype (Bennett et al. 2012). Lin et al. adopted two bi-level penalties in multitask regression model to integrate multiple diverse genomics datasets under different level and/or platform for identifying common biomarkers (e.g., genes or gene-set) (Lin et al. 2014a). They assumed a regression model for each dataset as a task, and then considered multiple regression models as multiple tasks. Variables from all datasets were grouped by specific units (e.g., genes) and penalized by sparse group penalties. The integration shows higher power of detecting risk genes than single omics data analysis and meta-analysis under the scenarios of both fixed effect and random effect.

It is noted that model-based integration methods need to build a model for each data set and then combine the models or their intermediate outputs. The scale of model errors or the intermediate outputs needs to be comparable for integration. If each omics data is extremely heterogeneous, this integration method may yield little improvement over separated analysis.

2.2.2.3 Latent Variable Models: Transform Variables into New Feature Space for Integration

The high dimensionality of diverse genomic data is a challenge. One commonly used strategy is to project high dimensional genomic data into low dimensional space before an integrative analysis is performed. Principle component analysis (PCA) is popularly used to explain the variance–covariance structure in a single data. It is widely used for handling pleiotropy with multiple correlated traits (e.g., eQTL) with the assumption that multiple correlated traits are able to reveal stronger signals than are obtained from univariate analysis of each trait separately. PCA based method collapses a number of correlated variables into a smaller number of uncorrelated variables as new phenotypes, which captures most variability and then test association for each new phenotype separately. Christine et al. used PCA to detect pleiotropic QTLs for boar taint and paternal fertility traits (Große-Brinkhaus et al. 2015). Jane et al. applied PCA on 70 skeletal traits to explore pleiotropy pattern through skeleton as well as genetic mechanism of each pattern (Kenney-Hunt et al. 2008).

Some latent variable models work in two- or multi-block way such as canonical correlation analysis (CCA) and partial least squares (PLS) with the aim to estimate latent variate from each dataset respectively (a linear combination of variables) by maximizing the correlation (CCA) or covariance (PLS) between them. Sonesson et al. applied CCA to explore two pairs of highly correlated features from the gene expression and copy number variable sets, which represent different characteristic

in leukemia. Tang et al. proposed a gene-based association test using CCA to detect QTLs associated with multiple quantitative traits (Tang and Ferreira 2012). Boulesteix et al. used PLS to predict transcription factor activities from combined analysis of gene expression and chromatin immunoprecipitation (ChIP) data (Boulesteix and Strimmer 2007). To integrate multiple datasets or clinical traits, some multi-block approaches such as multi-set CCA and multi-block PLS-correlation have also been proposed by summarizing pairwise correlations (or covariances) among different data sources (Lin et al. 2014b). In addition, parallel independent component analysis (pICA) and joint ICA are also two block methods widely used in genetic, imaging and clinical integration to explore independent components from each modality respectively while maximizing the correlation of the components simultaneously (Sui et al. 2012). Shen et al. show the robustness of joint ICA in integrating multi-omics data for biomarker detection and combined gene expression and copy number variation to identify significant genes associated with breast cancer (Sheng et al. 2011).

The above latent variables models mainly focus on the linear relationship among omics data. It may be interesting to consider non-linear relationship to explore more complicated genetic regulatory mechanism. ‘Kernel trick’ is a popular strategy which maps omics data into feature space by kernel matrix (e.g., Gaussian kernel matrix). Reverter et al. used kernel PCA to reduce dimension of metabolomics and genomics data and combined them for better representation of samples (Reverter et al. 2014). Yamannishi et al. proposed two types of kernel CCA to measure the correlation between several heterogeneous datasets, and to extract sets of genes which share similarities with respect to multiple biological attributes (Yamanishi et al. 2003).

Due to high dimensionality and small sample size of multi-omics data, there are usually issues of multi-collinearity (linear dependence) in the data and overfitting of the model. To address these issues, one way is to introduce the sparse regularizations into the conventional latent model to perform feature selection and correlative analysis simultaneously. Several types of regularized latent variable models have been proposed by enforcing different sparse penalties (e.g., lasso, elastic net and sparse group lasso penalty) on the loading vectors in the model. Waaijenborg et al. (2008) introduced the L-1 norm and elastic net penalties to the CCA model to analyze the correlation between gene expression and DNA-markers. Parkhomenko et al. (2009) proposed a CCA method with lasso penalty based on SVD (Singular value decomposition). Le Cao et al. (2009) used the penalized CCA with the elastic net to identify sets of co-expressed genes from two different microarray platforms. Witten et al. (2009) developed penalized matrix decomposition (PMD) method and applied it to solve CCA with lasso and fused lasso penalties. Lin et al. presented a unified framework of formulating these sparse CCA models as in (2.1):

$$\begin{aligned} \min_{\mathbf{u}, \mathbf{v}} -\mathbf{u}'\Sigma_{XY}\mathbf{v} + \lambda_1\|\mathbf{u}\|_G + \tau_1\|\mathbf{u}\|_1 + \lambda_2\|\mathbf{v}\|_G + \tau_2\|\mathbf{v}\|_1 \quad s.t. \mathbf{u}'\Sigma_{XX}\mathbf{u} \\ \leq 1, \mathbf{v}'\Sigma_{YY}\mathbf{v} \leq 1 \end{aligned} \quad (2.1)$$

where \mathbf{X}, \mathbf{Y} are the two data matrices; \mathbf{u} and \mathbf{v} are the loading vectors constrained by sparse terms; $\|\mathbf{u}\|_1$ and $\|\mathbf{v}\|_1$ are $l-1$ norm lasso penalty for performing the selection

of individual variable/feature, and $\|\mathbf{u}\|_G = \sum_{l=1}^L \omega_l \|\mathbf{u}_l\|_2$, $\|\mathbf{v}\|_G = \sum_{h=1}^H \mu_h \|\mathbf{v}_h\|_2$ are the group penalties to account for joint effects of features within the same group. The group penalty uses the non-differentiability of $\|\mathbf{u}_l\|_2$ (or $\|\mathbf{v}_h\|_2$) at $\mathbf{u}_l = 0$ ($\mathbf{v}_h = 0$) to set the coefficients of the group to 0 so the entire group of features will be removed to achieve the group sparsity.

Figure 2.1a shows the results of recovered loading vectors \mathbf{u} and \mathbf{v} by CCA-11, CCA-group and CCA-sparse group methods respectively. It can be seen that the CCA-sparse group method can better estimate true \mathbf{u} and \mathbf{v} than CCA-11, CCA-group method. Figure 2.1b compares the accuracy of recovering loading vectors from three methods with respect to different noise levels (standard deviation changes from 0.1 to 1 with interval 0.1), corresponding to different degrees of correlations between the two data sets. The result shows that the CCA-group model can recover the most correlated variables but gives the highest total discordance. CCA-sparse group has a comparable recovering accuracy as CCA-group model but much less total discordance especially when noise level decreases. These methods were also applied to fMRI data and SNP data and other omics data to identify significant correlated features.

Several other latent variable models were also proposed. Chun et al. proposed sparse PLS for simultaneous dimension reduction and feature selection in gene expression and transcriptional factor data. sPLS discriminant analysis (sPLS-DA), included in mixomics packages (Lê Cao et al. 2011), incorporated disease phenotype to extract those latent variables from gene expression or SNPs which are discriminative in multiclass disease, e.g., Leukemia. Li et al. introduced a sparse Multi-Block Partial Least Squares (sMBPLS) regression method to identify multidimensional regulatory modules from copy number variation, DNA methylation, gene expression and microRNA expression (Li et al. 2012).

2.2.2.4 Integrative Network Analysis

Networks represent the interactions of features within or across different levels of omics. The methods for reconstructing genetic network in single omics data have been well studied, as introduced in Sect. 2.2.1.4. However, they are limited to understand complex biological networks underlying cell and organ functions by single level of omic data. Integration of different levels of omics data to reconstruct comprehensive network is able to enrich our understanding of biological processes and improve the discovery of disease biomarkers. There are mainly two categories of integrative network reconstruction algorithms: single-stage reconstruction and multi-stage reconstruction.

Single-Stage Integrative Network Reconstruction This type of method tends to incorporate multi-omics data directly into the model for network construction. A simple way is using correlation based measurement to weight the interactions among omics features. WGCNA was used to construct network between metabolomics and transcriptomics data to identify clusters of metabolites and transcriptional factors associated with body weight change. A correlation derived

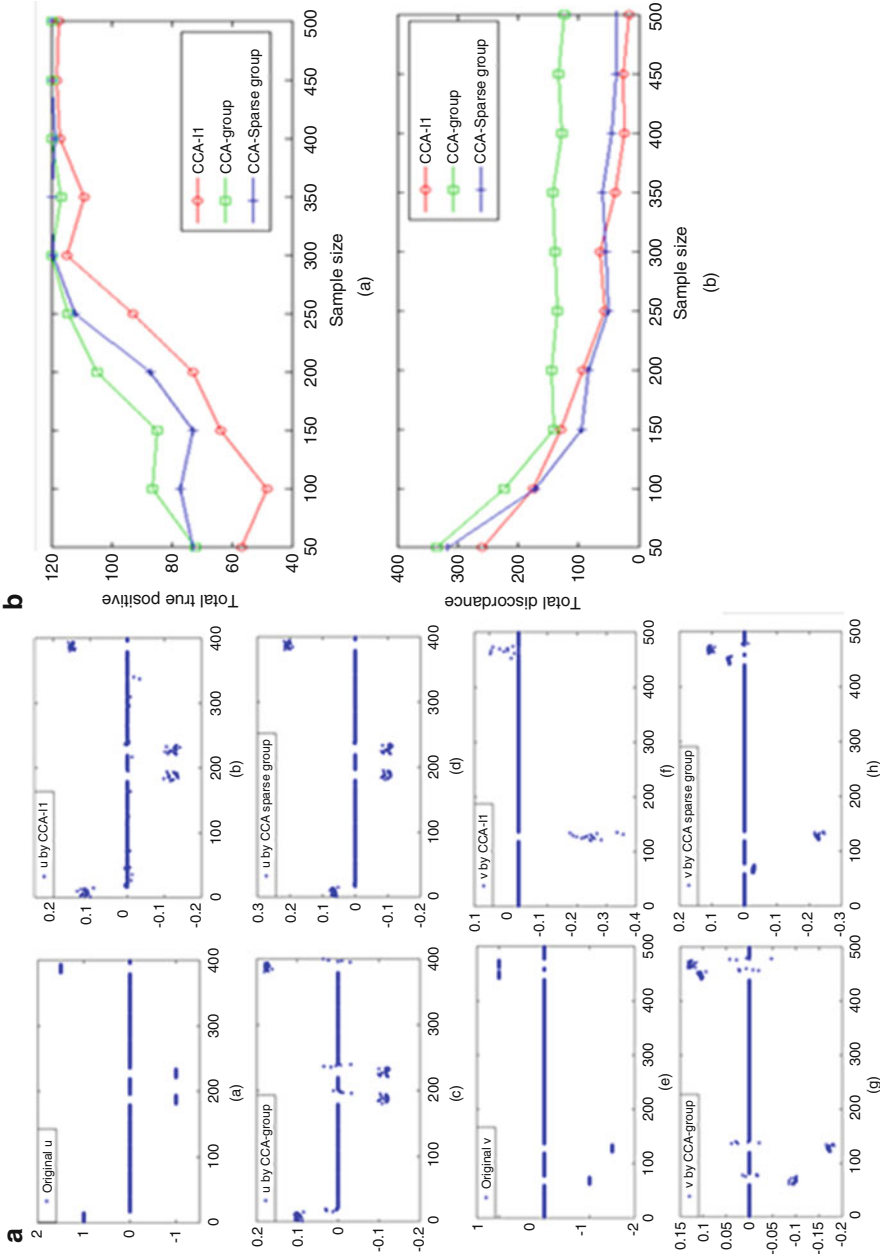


Fig. 2.1 A comparison of group sparse CCA with the other sparse CCA methods(e.g., CCA-lasso, CCA-group lasso). (a) The comparison of accuracy of recovering the loading vectors (u and v) by three methods. (b) The comparison of accuracy of recovering u and v total discordance with respect to different correlation values between omics datasets

topological matrix was used for clustering correlated features and cutting into different modules for association analysis (Wahl et al. 2015). Kayano et al. developed a statistical method based on low-order partial correlations with a robust correlation coefficient for estimating metabolic networks from metabolome, proteome, and transcriptome data (Kayano et al. 2013).

Another way is Bayesian network, which is a directed probabilistic graphical model with each edge representing the dependence between nodes (e.g., genes). Bayesian network is based on both prior distribution assumptions and observed data to design a model which can be mostly trusted. Prior distributions could be informative, such as conjugate prior, or mostly be non-informative. Some prior knowledge such as protein-protein interaction database could be incorporated to improve the accuracy and efficiency of network reconstruction. Conditional independence facilitates the integration of diverse data in a coherent way. Zhu et al. combined genotypic, expression, transcription factor binding site (TFBS), and protein-protein interaction (PPI) data to reconstruct causal gene networks. Three levels of Bayesian networks (BN_raw, BN_eQTL and BN_full) incorporating different prior knowledge (e.g., eQTL) were reconstructed and compared in terms of their power to infer causal regulators for validated signature gene sets (Zhu et al. 2008). Some Bayesian clustering models were designed to cluster genes from multiple omics data based on their interactions. Multiple dataset integration (MDI) was developed to identify groups of genes that are co-regulated and additionally their protein products appearing in the same complex (Kirk et al. 2012). To constrain the consistency of identified clusters across multiple omics sources, Bayesian consensus clustering was built to find consensus genetic clusters shared in different omics levels (Lock and Dunson 2013). Instead of finding clusters of genetic markers, Pathway recognition algorithm using data integration on genomic models (PARADIGM) was used to infer the molecular pathways altered in a patient sample by integrating genomic and functional genomic datasets (Vaske et al. 2010). Pathways were constructed based on prior knowledge database following CNV->gene expression->protein activity assumption and all measurements were categorized into three discrete states (inhibited, normal and activated). Joint posterior distribution was then computed based on observed data. The difference between pre- and post-activity levels indicated the quantitative alternation induced by the disease. Similarly, Multi-level Ontology Analysis (MONA) was a computationally efficient method to approximate the marginal posteriors of ontology terms based on three basis model assumptions (base, cooperative, and inhibitory models), given lists of genes responding to experimental conditions (Sass et al. 2013). iNET takes a “feature-specific” approach to model eight underlying biological basis models for constructing Bayesian network (Wang et al. 2013).

Multi-Stage Integrative Network Reconstruction There are generally two major steps: constructing network in each single level of omics data; and fusing multiple networks to an integrated network. The first step could be achieved by using various single omics network reconstruction algorithms. Network alignment and fusion methods are usually needed for the second step. Network alignment is the algorithm to map the nodes from two or multiple types of networks in such a way that

maximizes the topological and biological similarity between pairs of aligned nodes (Mitra et al. 2013). This technique is helpful in identifying previously undiscovered conserved modules that have been maintained across different species and revealing functionally similar subnetworks. Computational methods for network alignment consist of pair-wise alignment for aligning two networks only and multiple alignment to find transitive alignments among multiple networks. Some alignment algorithms, e.g., local alignment, aim to identify conserved regions between the input networks, which is particularly useful in finding known functional components (e.g., pathways) in a new species. For instance, PathBLAST allows the comparison of simple pathways (e.g., linear pathways) or subnetworks (e.g., modules) based on homology and interaction confidence (Kelley et al. 2004). NetworkBLAST finds highly conserved local regions greedily using inferred phylogeny (Kalaev et al. 2008). Some algorithms, e.g., global alignment, align every node in the smaller network to the larger network to find an overall network which enables species-level comparisons and discovery of functional orthologs. For instance, IsoRank and IsoRankN identify a stationary random walk distribution to perform global network alignment (Singh et al. 2008; Liao et al. 2009).

Network fusion is a technique to fuse multiple distinct but complementary biological networks to gain comprehensive insights of cellular structure and function. One of these approaches is integrating biological networks across different types of molecular interactions to identify composite modules. A cytoscape-based tool, PanGIA is designed to detect composite modules by identifying overlapping clusters of physical and genetic networks (Srivastava et al. 2011). Physical interactions are mainly represented by protein–protein and protein–DNA interactions. Genetic interactions represent functional relationships between genes, in which the phenotypic effect of one gene is modified by another. Composite modules are extracted based on the physical interactions while cluster of genetic interactions between two different composite modules reflect inter-modular dependencies. Integrative analysis of both physical and genetic networks can reveal physical mechanism of phenotype associated with genes in the composite module and also predict the genetic dependence between composite modules mapped in physical binding assays. Another Cytoscape tool, GeneMANIA builds a composite functional association network by taking a weighted average of individual functional association networks (Mostafaei et al. 2008). It first assigns weights to each of interaction networks. The composite network is then set to be the weighted average of the individual networks. Each network weights are calculated on demand and are tailored to the query list.

2.2.3 Statistics for Clinical Disease Diagnosis and Classification

The above has discussed the analysis of single omics or multi-omics data for biomarker detection, genetic regulatory network inferring as well as the exploration of genetic pathways underlying complex diseases. The next step is translating this

knowledge into clinical diagnosis or prediction. Predictive modeling, particularly classification, is critical in clinic research where risk biomarkers may vary largely with different diseases and even the subjects from one group may have subject-specific genetic variations. An effective method for classification of complex disease is demanded. We generally categorize them into two types: supervised learning method and unsupervised learning method. The former usually needs labelled training dataset for searching the optimal values of model parameters, which helps to build an accurate model and is more applicable for disease classification. The latter is data-driven method without knowing the class label from training, which is more likely to be used for subtyping to explore new subclass of diseases.

2.2.3.1 Supervised Learning in Omics Data

We will introduce several commonly used supervised classifiers in genetic data for classification of complex diseases. Assume there are m types of omics dataset, denoted by $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m]$, where $\mathbf{X}_i \in \mathbb{R}^{N \times P_i}$, $i = 1, 2, \dots, m$, P_i is the dimension of features in the i -th omics data. $\mathbf{Y} \in \mathbb{R}^{N \times c}$, c is the number of classes, and the subjects belonged to the j -th class are denoted by $\{w_j\}$, $j = 1, 2, \dots, c$. The object is to predict the class of a new sample \mathbf{y} given the observed omic feature matrices \mathbf{X} .

Discriminant Analysis Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are popularly used methods in clinical genomic analysis for risk feature identification and classification. LDA is a latent variable model which projects original high dimensional variables (e.g., gene expression measurements) into a new feature space by linear combinations $\mathbf{X}\boldsymbol{\alpha}$ with large ratios of between-group to within-group sums of squares, that is, maximizing the ratio $\boldsymbol{\alpha}^T \mathbf{B} \boldsymbol{\alpha} / \boldsymbol{\alpha}^T \mathbf{W} \boldsymbol{\alpha}$, where \mathbf{B} denotes the between-classes covariance matrix, and \mathbf{W} denotes the within-class covariance matrix. The calculation of \mathbf{B} and \mathbf{W} are given by

$$\mathbf{B} = \sum_{i=1}^C N(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T; \mathbf{W} = \sum_{i=1}^C \sum_{\mathbf{x} \in w_j} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T$$

where $\boldsymbol{\mu}_i = \frac{1}{N} \sum_{\mathbf{x} \in w_j} \mathbf{x}$, $\boldsymbol{\mu} = \frac{1}{N} \sum_{\forall \mathbf{x}} \mathbf{x}$. For a new subject \mathbf{x} , it can be projected to new

feature space by the estimated $\boldsymbol{\alpha}$ and classified to the class which has the minimum distance by the classification rule:

$$C(\mathbf{x}, L) = \operatorname{argmin}_k D_k(\mathbf{x})$$

where L is the training dataset to estimate LDA model and $D(\cdot)$ is the function to measure the distance between new subject with each class. LDA is a non-parametric method that is also a special form of a maximum likelihood

discriminant rule for multivariate normal class densities with the same covariance matrix. QDA is similar to LDA with the slight difference that QDA needs to estimate the covariance for each class separately. Zhang compared the two methods in recognition of two splice sites (acceptor site and donor site) in exons (Zhang 1997). The features from internal exons and their flanking regions (e.g., in-frame hexamer frequency bias) were adopted in LDA to distinguish acceptor site from donor site. To further consider the complex correlation structure among various acceptor sites or donor sites among exons, the covariance matrix may not be same between two sites. QDA was applied and shown better identification accuracy than LDA. There are also some other modifications of LDA to account for the specific characteristics in the omics data. For example, sparse LDA is combined with sparse regularizations to perform feature selection in discriminant analysis with high dimensional dataset, e.g., gene expression data (Clemmensen et al. 2011). Ye et al. also proposed unrelated LDA to handle the under-sampled data in genetic analysis and used generalized singular value decomposition method to make the features in transformed space be uncorrelated (Ye 2005). The method shows effectiveness in classification of tumors by gene expression data. Huang et al. compared LDA with other four modified methods on tumor classification by gene expression and showed the advantage of LDA modification methods over traditional LDA in terms of the average error and found no significant difference (Huang et al. 2009).

Decision Tree

Decision tree is one of most widely used machine learning methods. A decision tree model is built by a tree-like structure, where each internal node represents a specific test of an attribute, each branch represents one of the possible test results, and each leaf node represents an outcome. There are mainly two types of decision tree: decision tree classification and decision tree regression. The former aims to output the classifications labels (e.g., class) while the latter can output any real number of measurement. Decision tree can be learned by splitting the node into subsets according to the attribute value test. The splitting process is repeated in a recursive manner until the subsets of a node have all the same value of target variable or no more information could be added after splitting. Several algorithms have been developed to determine if splitting the node at each step, such as Gini impurity, information gain or variance reduction, leads to several types of decision trees, e.g., C4.5, C5, IDE, GINI, Codrington's and CART (classification and regression tree). Chen et al. used CART tree to select important genes for improving cancer classification (Chen et al. 2014). CART was also applied to explore the influence of the interactions among those genes that influence androgen in prostate cancer and if these interactions are able to improve the cancer prediction (Barnholtz-Sloan et al. 2011). There are also many other successful biological applications of decision tree based classification, including coding and noncoding DNA classification (Langfelder and Horvath 2008), protein secondary structure prediction (Shannon et al. 2003), and operon structure classification (Dennis et al. 2003).

Support Vector Machine

Support vector machines (SVM) are a family of classifiers which transform the input samples into a high dimensional space by a linear or kernel function, named feature space. Then a linear hyperplane could be drawn to separate two classes mapped in the feature space. To avoid overfitting, SVMs choose a specific hyperplane that maximizes the minimum distance from the hyperplane to the closest training point which is called support vectors. The optimal hyperplane is defined by the pair (w, b) by solving the following problem:

$$\begin{aligned} \min \quad & \|w\|^2 \\ \text{s.t.} \quad & y_i(w \cdot x_i + b) - 1 \geq 0, \forall i = 1, 2, \dots, N \end{aligned}$$

where $\|w\|^2$ measures the inverse of distance between two boundaries to obtain the maximum margin. $w \cdot x_i + b = \pm 1$ indicates two boundary hyperplanes separating subjects from two different classes ($y = 1$ or -1). Boundary hyperplanes are built on the support vectors. It is efficient for SVM to classify new examples since the majority of the training examples can be safely ignored. In order to transform original variables into high dimensional feature space and measure the non-linear correlation in feature space, a kernel function $K(x_i, x_j)$ is usually applied such as polynomial kernel, Gaussian radial basis function and hyperbolic function.

Support vector machines have drawn a lot of research efforts from diverse fields (Noble 2004). In bioinformatics, it is widely used for cancer diagnosis and classification, protein structure and function prediction and gene expression pattern recognition. An early application example of SVM is to identify important genes and further improve the classification on leukemia and colon cancers (Guyon et al. 2002). Ferry et al. used SVM to not only classify cancer tissue samples based on microarray data but also identify those samples wrongly classified by experts. Hua and Sun used SVMs to perform protein classification with respect to subcellular localization (Hua and Sun 2001). A 20-feature composition kernel function is applied and shown to produce more accurate classifications than other competing methods, including a neural network, a Markov Distinguishing model and the covariant discriminant algorithm. Yeang et al. extended SVM to multi-class SVM which can address the multiple classes issue. The method was applied for multi-class tumor classification on a data set of 190 samples from 14 tumor classes (Yeang et al. 2001). Nguyue et al. compared several multi-lass SVM algorithms on protein secondary structure prediction including: one-against-all, one-against-one, and directed acyclic graph, and two approaches for multi-class problem by solving one single optimization problem (Nguyen and Rajapakse 2003). The results demonstrated better recovery accuracy of multi-class SVMs proposed by Vapnik and Weston than the other multi-class SVMs, including binary SVMs.

Ensemble Learning

Ensemble learning is an effective technique that constructs a set of classifiers and combines them to improve overall prediction accuracy (Dietterich 2000). There are a lot of ensemble methods that have been applied to biological data analysis in addressing small sample size but high dimensional data sets and reducing the overfitting risk. The classification accuracy is also improved by generating multiple prediction models and aggregating these multiple models (called basis classifiers) to make the final prediction in a consensus way. There are several types of ensemble learning algorithms including bagging (Breiman 1996), boosting (Freund and Schapire 1996) and random forests (Breiman 2001). Being the principle ensemble learning methods, they are usually combined with the other classifiers such as decision trees.

There are several applications of ensemble learning methods such as sample/tissue classification and gene-gene interaction prediction. Ben-Dor et al. (2000) and Dudoit et al. (2002) applied bagging and boosting methods to classify tumors using gene expression profiles. Both studies compared the ensemble methods with other individual classifiers such as k-nearest neighbors (kNN), clustering based classifiers, SVM, LDA, and classification trees. The conclusion was that ensemble methods (e.g., bagging and boosting) performed similarly to other single classification algorithms. Wu et al. (2003), compared several methods for the classification of ovarian cancer based on MS spectra including the ensemble methods of bagging, boosting, and random forests to individual classifiers, e.g., LDA, QDA, kNN, and SVM. The study found that among all methods random forests outperforms the others with the lowest error rate. Moon et al. developed a new ensemble-based classification algorithm, Classification by Ensembles from Random Partitions (CERP) combined with classification and decision tree (CART) and applied it to genomic data on leukemia patients and on breast cancer patients (Moon et al. 2006). The performance was compared with other classifiers such as single decision tree (e.g., CART), SVM, diagonal LDA and other ensemble learning methods (e.g., RF and boosting). The results demonstrate that CERP is a consistently better algorithm and maintains a good balance between sensitivity and specificity even in case of unbalanced sample size.

2.2.3.2 Unsupervised Learning in Omics Data

Clustering is a popular unsupervised learning method and commonly applied in omics data analysis such as clustering genes based on their expression, or clustering samples based on their omics features to identify subgroups or subtypes of diseases. There are several clustering methods proposed including partition clustering and hierarchical clustering.

Partition Clustering

This type of clustering methods mainly partition objects and change the clusters based on the dissimilarity or distance between objects with clusters. The fixed number of clusters could be specified before the clustering.

K-means clustering is a popular method for clustering genes or subjects. The general procedure is as follows:

- (1) Randomly generate k clusters and calculate the centroid of each cluster;
- (2) Calculate the distance of each point with each cluster centroid and assign each point to the cluster with shortest distance.
- (3) Update the centroid of each new cluster;
- (4) Repeat until certain convergence is met, e.g., no changes of assignment of each point.

There are some applications of k-means in bioinformatics, such as gene clustering or subtyping. Lehmann et al. used k-means to analyze gene expression profiles of 587 TNBC cases from 21 breast cancer to subtype TNBC. Each TNBC case contained 13,060 genes after normalization for clustering analysis by K-means. The optimal number of clusters was determined by the change of proportion of area under empirical cumulative distribution curve and consequently, 6 Triple-negative breast cancer subtypes were identified with unique gene expression and ontologies (Lehmann et al. 2011). Further they predicted “driver” signaling pathways of each subtypes to show that analysis of distinct GE signatures can inform therapy selection.

Fuzzy C-means (FCM) clustering is another clustering method using the ‘soft’ clustering instead of ‘hard’ clustering in k-means. For each subject, FCM assigns a degree of membership in each cluster, which can account for the uncertainty of some subjects. It has been widely used in imaging analysis (Li et al. 2013) since it is more suitable for the scenario that there is overlapping among clusters, which is also common in clinical analysis such as tumor classification where unlabeled tumor samples may not necessarily be clear members of one class or another. Wang et al. applied FCM clustering on gene expression data for tumor classification and gene prediction (Wang et al. 2003). Given a dataset $X = [X_1, X_2, \dots, X_N] \in R^{N \times p}$ from N tumor subjects measured on p gene expression levels. We assume the existence of N_c tumor classes, whose centers are denoted by $C = [C_1, C_2, \dots, C_{N_c}]$ which are unknown and to be estimated. $U = [U_{i,1}, U_{i,2}, \dots, U_{i,N_c}]$ is fuzzy membership matrix for the i -th subject on all of tumor classes, whose value between zero and one. FCM clustering can be obtained by solving the optimization issue:

$$\min_{U,C} \sum_{k=1}^{N_c} \sum_{i=1}^N u_{k,i}^q \|X_i - C_k\|^2, \text{ subject to } \sum_{k=1}^{N_c} u_{k,i}^q = 1$$

where q is a weight on each fuzzy membership and determines the degree of fuzziness. Each tumor subject will have a membership in every class; membership

close to one indicates a high degree of similarity between the subject and a tumor class while membership close to zero implies little similarity. The subject is assigned to the class with the highest membership values. The second term is used to constrain that the summation of membership of different classes equals one to make sure the value of membership is between zero and one. The tests on four different tumor datasets show the efficiency of FCM clustering in terms of reduced error rates and the importance of selected features for medical diagnostics and cancer classification.

Hierarchical Clustering

Hierarchical clustering is a clustering method to represent the objects in a tree-like structure, where each node has zero or more child nodes below it. There are mainly two types of strategies to generate the hierarchical tree: agglomerative, a ‘bottom up’ approach which takes each object as its own cluster and merge clusters as one moves up the hierarchy; divisive, a ‘top down’ approach which takes all objects as one cluster and split it recursively as one moves down the hierarchy. Here shows the procedure of agglomerative as an example:

- (1) Start with n clusters with each contains one object;
- (2) Merge the most similar pair of clusters from the proximity matrix which can be built based on different distance measurements, e.g., single linkage, complete linkage and average linkage, which take the minimum, maximum and average of pairwise distance between two clusters, respectively.
- (3) Update the proximity matrix by replacing the individual clusters with merged cluster;
- (4) Repeat until only one cluster is left.

Hierarchical clustering is also applied for clinical classification and gene clustering. Makretsov et al. used hierarchical clustering to determine the efficiency in improving prognostication in patients with invasive breast cancer by multiple immunomarkers (protein expression profiles) (Makretsov et al. 2004). They identified three cluster groups with significant differences in clinical outcome and demonstrated that hierarchical clustering by using multiple markers can group breast cancers into classes with clinical relevance and outperform individual prognostic markers. Furlan et al. applied unsupervised hierarchical clustering analysis to 126 colorectal carcinomas to combine 13 routinely assessed clinicopathologic features and all five molecular markers to distinguish four molecular subtypes of sporadic colorectal carcinomas (Furlan et al. 2011). The results demonstrate the superiority of classification based on the combination of clinicopathologic and molecular features of colorectal cancers over single features, and also indicate that hierarchical clustering is a useful tool to define a diagnostic and prognostic signature for different carcinomas.

References

- Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing statistical methods for constructing large scale gene networks. *PLoS ONE*. 2012;7:e29348.
- Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet*. 2006;7:781–91.
- Barnholtz-Sloan JS, Guan X, Zeigler-Johnson C, Meropol NJ, Rebbeck TR. Decision tree-based modeling of androgen pathway genes and prostate cancer risk. *Cancer Epidemiol Biomark Prev*. 2011;20:1146–55.
- Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, et al. Tissue classification with gene expression profiles. *J Comput Biol*. 2000;7:559–83.
- Bennett BD, Xiong Q, Mukherjee S, Furey TS. A predictive framework for integrating disparate genomic data types using sample-specific gene set enrichment analysis and multi-task learning. *PLoS ONE*. 2012;7:e44635.
- Boulesteix A-L, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform*. 2007;8:32–44.
- Breiman L. Bagging predictors. *Mach Learn*. 1996;24:123–40.
- Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
- Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am J Hum Genet*. 2010;86:6–22.
- Chari R, Coe BP, Vucic EA, Lockwood WW, Lam WL. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Syst Biol*. 2010;4:67.
- Chen K-H, Wang K-J, Tsai M-L, Wang K-M, Adrian AM, et al. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. *BMC Bioinf*. 2014;15:49.
- Clemmensen L, Hastie T, Witten D, Ersbøll B. Sparse discriminant analysis. *Technometrics*. 2011;53:406–13.
- Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, et al. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*. 2003;4:P3.
- Dieterich TG. Ensemble methods in machine learning. In: *Multiple classifier systems*. Berlin/Heidelberg: Springer; 2000. p. 1–15.
- Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc*. 2002;97:77–87.
- Efron B, Tibshirani R. On testing the significance of sets of genes. *The Annals of Applied Statistics*. 2007;1:107–29.
- Ferreira MA, Purcell SM. A multivariate test of association. *Bioinformatics*. 2009;25:132–3.
- Freund Y, Schapire RE. Experiments with a new boosting algorithm. 1996:148–56.
- Fridley BL, Lund S, Jenkins GD, Wang L. A Bayesian integrative genomic model for pathway analysis of complex traits. *Genet Epidemiol*. 2012;36:352–9.
- Friedman N, Linial M, Nachman I, Pe’er D. Using Bayesian networks to analyze expression data. *J Comput Biol*. 2000;7:601–20.
- Furlan D, Carnevali IW, Bernasconi B, Sahnane N, Milani K, et al. Hierarchical clustering analysis of pathologic and molecular data identifies prognostically and biologically distinct groups of colorectal carcinomas. *Mod Pathol*. 2011;24:126–37.
- Galesloot TE, van Steen K, Kiemeny LA, Janss LL, Vermeulen SH. A comparison of multivariate genome-wide association methods. *PLoS ONE*. 2014;9:e95923.
- Große-Brinkhaus C, Storck LC, Frieden L, Neuhoﬀ C, Schellander K, et al. Genome-wide association analyses for boar taint components and testicular traits revealed regions having pleiotropic effects. *BMC Genet*. 2015;16:36.
- Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46:389–422.

- He H, Zhang L, Li J, Wang YP, Zhang JG, et al. Integrative analysis of GWASs, human protein interaction, and gene expression identified gene modules associated with BMDs. *J Clin Endocrinol Metab.* 2014;99:E2392–9.
- Holzinger ER, Ritchie MD. Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics.* 2012;13:213–22.
- Holzinger ER, Dudek SM, Frase AT, Pendergrass SA, Ritchie MD. ATHENA: the analysis tool for heritable and environmental network associations. *Bioinformatics.* 2013;30(5):698–705.
- Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics.* 2001;17:721–8.
- Huang D, Quan Y, He M, Zhou B. Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data. *J Exp Clin Cancer Res.* 2009;28:149.
- Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. *Stat Sci Rev J Instit Math Stat.* 2012;27:481–99.
- Jansen RC, Nap J-P. Genetical genomics: the added value from segregation. *TRENDS Genet.* 2001;17:388–91.
- Kalaev M, Smoot M, Ideker T, Sharan R. NetworkBLAST: comparative analysis of protein networks. *Bioinformatics.* 2008;24:594–6.
- Kayano M, Imoto S, Yamaguchi R, Miyano S. Multi-omics approach for estimating metabolic networks using low-order partial correlations. *J Comput Biol.* 2013;20:571–82.
- Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, et al. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res.* 2004;32:W83–8.
- Kendzioriski C, Chen M, Yuan M, Lan H, Attie A. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics.* 2006;62:19–27.
- Kenney-Hunt JP, Wang B, Norgard EA, Fawcett G, Falk D, et al. Pleiotropic patterns of quantitative trait loci for 70 murine skeletal traits. *Genetics.* 2008;178:2275–88.
- Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, et al. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat Genet.* 2008;40:904–8.
- Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol.* 2012;8:e1002375.
- Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics.* 2012;28:3290–7.
- Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet Epidemiol.* 2008;32:9–19.
- Krishnapuram B, Carin L, Figueiredo MA, Hartemink AJ. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transac Pattern Anal Mach Intell.* 2005;27:957–68.
- Lando M, Holden M, Bergersen LC, Svendsrud DH, Stokke T, et al. Gene dosage, expression, and ontology analysis identifies driver genes in the carcinogenesis and chemoradioresistance of cervical cancer. *PLoS Genet.* 2009;5:e1000719.
- Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* 2008;9:559.
- Le Cao KA, Martin PGP, Robert-Granie C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinf.* 2009;10:34.
- Lê Cao K-A, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinf.* 2011;12:253.
- Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest.* 2011;121:2750.
- Li W, Zhang S, Liu C-C, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics.* 2012;28:2458–66.
- Li J, Lin D, Cao H, Wang Y-P. An improved sparse representation model with structural information for Multicolour Fluorescence In-Situ Hybridization (M-FISH) image classification. *BMC Syst Biol.* 2013;7:S5.

- Liao C-S, Lu K, Baym M, Singh R, Berger B. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*. 2009;25:i253–8.
- Lin D, Zhang J, Li J, He H, Deng H-W, et al. Integrative analysis of multiple diverse omics datasets by sparse group multitask regression. *Frontiers in cell and developmental biology*. 2014a;2:62.
- Lin D, Cao H, Calhoun VD, Wang Y-P. Sparse models for correlative and integrative analysis of imaging and genetic data. *J Neurosci Methods*. 2014b;237:69–78.
- Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics*. 2013;29(20):2610–6.
- Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. *Ann Stat*. 2014;42:413.
- Louhimo R, Hautaniemi S. CNAmets: an R package for integrating copy number, methylation and expression data. *Bioinformatics*. 2011;27:887–8.
- Makretsov NA, Huntsman DG, Nielsen TO, Yorlida E, Peacock M, et al. Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma. *Clin Cancer Res*. 2004;10:6143–51.
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*. 2007;39:906–13.
- Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinf*. 2006;7 Suppl 1:S7.
- Melzer D, Perry JR, Hernandez D, Corsi A-M, Stevens K, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet*. 2008;4:e1000072.
- Mitra K, Carvunis A-R, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet*. 2013;14:719–32.
- Moon H, Ahn H, Kodell RL, Lin C-J, Baek S, et al. Classification methods for the development of genomic signatures from high-dimensional data. *Genome Biol*. 2006;7:R121.
- Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*. 2008;9:S4.
- Nguyen MN, Rajapakse JC. Multi-class support vector machines for protein secondary structure prediction. *Genome Inform*. 2003;14:218–27.
- Noble WS. Support vector machine applications in computational biology. In: *Kernel methods in computational biology*. The MIT Press; 2014. p. 71–92.
- Ogutu JO, Schulz-Streeck T, Piepho H-P. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BioMed Cent Ltd*. 2012;6(2):1–6.
- O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS ONE*. 2012;7:e34861.
- Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. *Stat Appl Genet Mol Biol*. 2009;8:1–34.
- Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. *J Am Stat Assoc*. 2009;104:735–46.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–75.
- Reverter F, Vegas E, Oller JM. Kernel-PCA data integration with enhanced interpretability. *BMC Syst Biol*. 2014;8:S6.
- Sass S, Buettner F, Mueller NS, Theis FJ. A modular framework for gene set analysis integrating multilevel omics data. *Nucleic Acids Res*. 2013;41:9622–33.
- Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*. 2005;37:710–17.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498–504.

- Sheng J, Deng H-W, Calhoun V, Wang Y-P. Integrated analysis of gene expression and copy number data on gene shaving using independent component analysis. *IEEE/ACM Transac Comput Biol Bioinform (TCBB)*. 2011;8:1568–79.
- Shevade SK, Keerthi SS. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*. 2003;19:2246–53.
- Singh R, Xu J, Berger B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc Natl Acad Sci*. 2008;105:12763–8.
- Soneson C, Lilljebjörn H, Fioretos T, Fontes M. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinf*. 2010;11:191.
- Srivastava R, Hannum G, Ruschinski J, Ono K, Wang P-L, et al. Assembling global maps of cellular function through integrative analysis of physical and genetic networks. *Nat Protoc*. 2011;6:1308–23.
- Stephens M. A unified framework for association analysis with multiple related phenotypes. *PLoS ONE*. 2013;8:e65245.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102:15545–50.
- Sui J, Adali T, Yu Q, Chen J, Calhoun VD. A review of multivariate methods for multimodal fusion of brain imaging data. *J Neurosci Methods*. 2012;204:68–81.
- Tang CS, Ferreira MA. A gene-based test of association using canonical correlation analysis. *Bioinformatics*. 2012;28:845–50.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc Ser B (Methodol)*. 1996;58(1):267–88.
- Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001;98:5116–21.
- Tyekucheva S, Marchionni L, Karchin R, Parmigiani G. Integrating diverse genomic data using gene sets. *Genome Biol*. 2011;12:R105.
- van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet*. 2013;9:e1003235.
- Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. 2010;26:i237–45.
- Waaijenborg S, Hamer PCVDW, Zwinderman AH. Quantifying the association between gene expressions and DNA-Markers by penalized canonical correlation analysis. *Stat Appl Genet Mol Biol*. 2008; 7
- Wahl S, Vogt S, Stückler F, Krumsiek J, Bartel J, et al. Multi-omic signature of body weight change: results from a population-based cohort study. *BMC Med*. 2015;13:48.
- Wang J, Bø TH, Jonassen I, Myklebost O, Hovig E. Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. *BMC Bioinf*. 2003;4:60.
- Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*. 2007;81:1278–83.
- Wang S, Nan B, Zhu N, Zhu J. Hierarchically penalized Cox regression with grouped variables. *Biometrika*. 2009;96:307–22.
- Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*. 2010;11:843–54.
- Wang W, Baladandayuthapani V, Holmes CC, Do K-A. Integrative network-based Bayesian analysis of diverse genomics data. *BMC Bioinf*. 2013;14:S8.
- Werhli AV, Grzegorzczak M, Husmeier D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*. 2006;22:2523–31.
- Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*. 2009;10:515–34.

- Wu B, Abbott T, Fishman D, McMurray W, Mor G, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*. 2003;19:1636–43.
- Xiong Q, Ancona N, Hauser ER, Mukherjee S, Furey TS. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Res*. 2012;22:386–97.
- Yamanishi Y, Vert J-P, Nakaya A, Kanehisa M. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*. 2003;19:i323–30.
- Ye J. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *J Mach Learn Res JMLR*. 2005;6:483–502.
- Yeang C-H, Ramaswamy S, Tamayo P, Mukherjee S, Rifkin RM, et al. Molecular classification of multiple tumor types. *Bioinformatics*. 2001;17:S316–22.
- Zhang MQ. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc Natl Acad Sci*. 1997;94:565–8.
- Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005;4:Article17.
- Zhu J, Zhang B, Smith EN, Drees B, Brem RB, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet*. 2008;40:854–61.

Application of Clinical Bioinformatics

Wang, X.; Baumgartner, C.; Shields, D.C.; Deng, H.-W.;
Beckmann, J.S. (Eds.)

2016, VI, 398 p. 120 illus., 70 illus. in color., Hardcover

ISBN: 978-94-017-7541-0