# Chapter 8
# Systems Analysis of High-Throughput Data

**Rosemary Braun**

**Abstract** Modern high-throughput assays yield detailed characterizations of the genomic, transcriptomic, and proteomic states of biological samples, enabling us to probe the molecular mechanisms that regulate hematopoiesis or give rise to hematological disorders. At the same time, the high dimensionality of the data and the complex nature of biological interaction networks present significant analytical challenges in identifying causal variations and modeling the underlying systems biology. In addition to identifying significantly disregulated genes and proteins, integrative analysis approaches that allow the investigation of these single genes within a functional context are required. This chapter presents a survey of current computational approaches for the statistical analysis of high-dimensional data and the development of systems-level models of cellular signaling and regulation. Specifically, we focus on multi-gene analysis methods and the integration of expression data with domain knowledge (such as biological pathways) and other gene-wise information (e.g., sequence or methylation data) to identify novel functional modules in the complex cellular interaction network.

**Keywords** Statistical analysis · High-throughput data · Microarrays · Sequencing · NGS · Genomics · Machine learning · Network models

## Introduction

The precise coordination of complex and adaptive living processes relies upon systems that regulate transcriptional, posttranscriptional, and epigenetic control of gene expression and protein production. In contrast to the simplified view of the "central dogma" of molecular biology, wherein transcription followed by translation leads linearly from DNA to RNA to protein, it is now understood that there exist feedback loops at each stage, forming a network of regulatory interactions (Fig. 8.1).

R. Braun (✉)
Biostatistics Division, Department of Preventive Medicine
and Northwestern Institute on Complex Systems, Northwestern University,
680 N. Lake Shore Dr., Suite 1400, Chicago, IL 60611, USA
e-mail: rbraun@northwestern.edu



S. J. Corey et al. (eds.), *A Systems Biology Approach to Blood,*
Advances in Experimental Medicine and Biology 844, DOI 10.1007/978-1-4939-2095-2_8

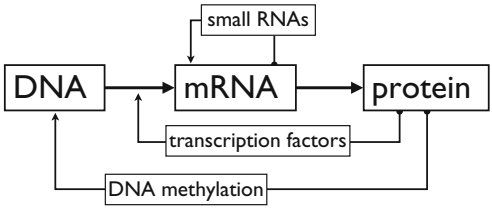


**Fig. 8.1** Regulatory mechanisms in molecular biology. DNA is transcribed to mRNA and then translated into protein. The rate of transcription is controlled by a feedback loop in which the level of transcription factor proteins is regulated the activity of the transcriptional complex, and genes can be permanently silenced by methylation of cytosine in CpG promoter regions of the DNA sequence. More recently, it has been discovered that the expression of small noncoding RNA molecules (e.g., microRNAs) can downregulate entire sets of genes by binding to complementary sequences in the mRNA

Identifying functionally relevant genes and unraveling the systems governing their expression can elucidate the molecular mechanisms underlying development and disease, as well as facilitate the development of prognostic tests and therapeutic interventions [1, 2].

Although living organisms have long been thought of as complex systems comprising many strongly interdependent parts [3, 4], the study of biological processes at the systems level remained a theoretical practice until fairly recently. Prior to the completion of the Human Genome Project and the development of high-throughput technologies, limitations on the ability to exhaustively assay samples of interest required that each gene be probed one at a time, leading to a reductionist approach in which biological systems were investigated by examining their parts in isolation. In recent years, however, major technological advances have enabled assays that yield highly detailed genome-wide information for each sample (including sequence, expression, and epigenetic modifications). This unprecedented increase in our ability to probe how every gene is expressed in a particular tissue or responds to a particular environmental perturbation now makes systems biology possible. The wealth of data now being generated in high-throughput profiling studies not only allows gene-level analyses to be applied comprehensively across the entire genome, but provides an immense opportunity to augment reductionist one-gene-at-a-time techniques with systems-level analyses that treat the data in an integrative manner and elucidate the functional association between differentially expressed genes.

Complementing the advances in experimental technologies, advances in computing technology have ushered in an exciting era of computational systems biology. Broadly speaking, computational systems biology investigations may be classified into two groups, each with its own utility and set of challenges: the statistical analysis of high-dimensional data to infer differentially regulated network modules from experimental studies, and the dynamical simulation of these networks to model the occurrence of cellular events. Here, we focus on statistical and machine learning algorithms to draw inferences about regulatory networks from complex data sets. Combined with gene-level analyses, pathway-based methods provide comprehensive

analyses of the functional modules that govern biological processes. The objective of this chapter is to provide theoretical and practical knowledge of how high-throughput data can be harnessed to yield mechanistic insights and build predictive models at the systems level.

## Generating High-Throughput Data

The accuracy of any systems-level analysis will depend on the quality of the data being analyzed. This, in turn, depends upon the experimental design, the assay technology employed, and the preprocessing of the raw data. Although a full review of these considerations is beyond the scope of this chapter, a brief overview is presented for context.

### *Experimental Design*

Experiments may be designed with several goals in mind:

*Class comparison* Identification of genes or gene sets behaving differently between predetermined "classes" of samples (e.g., cases and controls, different phenotypes, different stages of development, different treatments, etc.).

*Time series* Investigation of the dynamics of gene expression changes following an exposure (e.g., to examine how the expression profile changes over time and differs between growth phenotypes).

*Class prediction (supervised machine learning)* Identification of a minimal set of genes that can be used to categorize a new sample into one of several known types based on its molecular profile (e.g., with the goal of predicting treatment response). Also called supervised machine learning.

*Class discovery (clustering/unsupervised ML)* Identification of novel groups of samples on the basis of their molecular profiles (e.g., to identify disease subtypes among clinically similar cases that may correspond to differing prognoses).

*Network Analysis* Identification of differential relationships between molecules, either by analyzing the data in the context of putative interaction networks or by "reverse engineering" the underlying network based on experimental data.

Regardless of the question under consideration, several guiding principles should be observed. First, all high-throughput studies yield a measurements in a feature space ($10^5$–$10^6$ probes) that is of much higher dimensionality than the number of samples (often on the order of $10^2$). From a mathematical modeling standpoint, these experiments are underdetermined, meaning there are many more variables (genes) than there are equations (samples), and different analysis methods may yield different results that are nevertheless equally valid/optimal fits. Second, despite improvements in quality control and experimental accuracy and precision, high-throughput

technologies remain relatively noisy and are highly sensitive to batch effects (meaning that the same samples, assayed at two different labs or at two different times using identical protocols, may exhibit highly differentially expressed genes that are responding to extraneous biological variables). These two challenges underscore the need for biological replicates: both to increase the power of the many gene-wise statistical tests being performed, and to capture the natural level of variability between phenotypically identical samples.

## *Microarrays*

There currently exist a number of different experimental modalities for genomic investigations, each with its own benefits and challenges. The oldest and best-established are microarrays, which measure the hybridization of fluorophore-labeled nucleic acid strands to complementary probe sequences on a chip. The intensity of fluorescence at a specific probe spot is proportional to the amount of bound nucleic acid strands. Microarray chips contain $10^5$–$10^6$ different probes, permitting thousands of genes to be simultaneously assayed. These may be designed to measure mRNA abundance (gene expression profiling), microRNAs (miRNA profiling), or to detect single nucleotide polymorphisms (SNPs) in DNA. Chips functionalized with antibodies may be used in a similar fashion to assess protein abundance.

Before they can be analyzed, microarray data must be preprocessed and normalized. The preprocessing steps include the subtraction of background intensities, averaging across duplicated probes, thresholding or scaling to spiked-in controls or housekeeping genes, removal of probes that fail to meet QC criteria, and normalization to render each array comparable to the others. Normalization schemes rely upon the assumption that the vast majority of genes are not differentially modulated in the phenotype of interest, and attempt to remove chip-wide variations in gene expression that are likely due to technical factors alone. The choice of preprocessing and normalization algorithms can have a significant impact on the results of the statistical analysis, and the appropriate selection depends in part on the microarray technology; the reader is referred to the several comprehensive reviews [5–7] for additional guidance. Because the normalized abundances are approximately log-normally distributed, values expressed on a logarithmic scale are often tested using standard parametric statistics.

## *"Next Generation" Sequencing*

The development of next generation sequencing (NGS) represents an important leap forward in identifying disease-specific genetic variants (DNAseq), epigenetic modifications (ChIPSeq of histone methylation), and transcriptional regulation and splicing (RNAseq). Combined, such genomic data provide a powerful means to identify the relationships between the genetic sequence, epigenetic marks, and expression of genes.

In contrast to microarrays, which probe regions of the genome with known sequences, NGS studies comprehensively assay the entire genome. The data produced are vast, and present different preprocessing challenges than those encountered in microarray studies. The experimental technique consists of fragmenting the DNA or RNA into short segments, which are then sequenced. These so-called "short reads" must then be aligned to a reference genome sequence in order to identify the genes to which they correspond. (Although NGS assays are highly comprehensive, the mapping of reads is a computationally challenging task, and the resulting data is often considerably noisier than that obtained by microarray.) The number of reads for a given genomic region is used as a measure of gene expression (in RNAseq) and to identify probable transcription-factor binding sites or epigenetic modifications (DNAseq, ChIPseq). For more details on sequencing, alignment, and variant calling in NGS studies, the reader is referred to two recent reviews [8, 9]. Once these steps are completed, the data may be analyzed to reveal disease-associated genetic variants, epigenetic modifications, and differential expression [10].

## *Gene-Level Statistical Analyses*

While the focus of this chapter is to acquaint the reader with systems-level statistical analyses, it is useful to briefly review several common gene-level approaches.

Often, the first goal is to identify genes that behave differently in the sample groups of interest ("class comparison"). For mRNA and miRNA expression studies, where the gene level data are continuous, genes are tested for differential expression between groups using a $t$-test; where more covariates are involved (such as in studies investigating gene$\times$environment interactions), linear models may be used. Linear models may also be used in the context of time-series analysis to identify genes whose expression changes over time and detect those whose time-course profiles differ between sample classes. In SNP and sequence studies, where the covariates are categorical, $\chi^2$ tests are used to identify SNP loci where minor allele frequencies differ significantly. These tests yield a statistic, one per gene/miRNA/locus, that quantifies the difference in expression or allele frequencies between the groups of interest. These statistics may then be compared against an appropriate distribution to yield a $p$-value and identify significant associations. (For expression microarrays, the `limma` package [11] in R [12] provides a user-friendly framework for gene level analyses. Other BioConductor utilities [13, 14] provide similar functionality for SNP arrays, NGS, and other experimental modalities.)

In all cases, the vast number of hypotheses being tested (at least one per gene, and often times more) necessitates a multiple testing correction [15] of the $p$-values. That is, at a significance threshold of $p \leq 0.05$, we expect that a gene will be falsely called significant 5 % of the time, leading to thousands of such false positives when the number of genes assayed is on the order of $10^5$. While the simple Bonferroni adjustment may be used (in which the significance threshold is set to 0.05 divided by the number of genes assayed), it is considered to be excessively conservative.

Specifically, it is assumed in the Bonferroni adjustment that each gene is strictly independent of all others, an assumption well known to be false for genomic data (in the case of expression, co-regulated genes will exhibit correlated expression; in SNPs, patterns of recombination will lead to linkage disequilibrium, or a tendency for SNP alleles at one loci to be correlated with the alleles at another). Instead, the false positives should be controlled using the false discovery rate adjustment (FDR) [16], which has been proven to exert robust control over the error rate even when the hypotheses have dependencies [17]. Alternatively, assumption-free but computationally intensive permutation procedures [18] may be used.

## Identifying Functional Modules

The lists of significant genes obtained by the analyses described above provide limited mechanistic insights without additional biological context. To gain an understanding of systems biology, it is necessary to assemble single-gene information to identify sets of genes and interactions that fulfill particular biological functions. Typically, this is done either by finding clusters of genes that behave in the same way in the experiment, or by incorporating expert knowledge from pathway databases to focus the analysis.

### *Clustering*

It is well-accepted that genes interact with each other in transcriptional modules, and that these modules in turn interact with other modules [19, 20]. Because of these relationships, genes that function together often exhibit directly or inversely correlated expression. The simplest method for identifying those modules and connections is by clustering the genes to identify groups of genes whose expression is similar across the set of samples [21, 22].

The two most commonly used clustering algorithms are hierarchical clustering and $k$-means clustering. Their considerable popularity is due to their computational and conceptual simplicity. However, because both rely upon the user to specify the number of clusters, they are prone to artificially separating genes that should be in the same cluster (if the user specifies more clusters than are truly present) or speciously combining them (if the user specifies too few). They are limited in their ability to detect clusters with complex shapes. To address these limitations, refinements of both schemes have been proposed; we describe them below.

#### Hierarchical Clustering

The commonly used hierarchical clustering [23] technique agglomeratively sorts genes based on the similarity of their expression, producing a tree that can be cut

into clusters. For each pair of genes $i$ and $j$, hierarchical clustering computes a distance metric $D_{ij}$; then, starting with each gene as its own "cluster," iteratively merges clusters based on the smallest $D_{ij}$ between them. Most frequently, a Euclidean distance metric (i.e., $D_{ij}=\sqrt{\sum_m (g_{i,m}^2+g_{j,m}^2)}$ where $g_{i,m}$ denotes the expression of gene $i$ in sample $m$) is used, although non-Euclidean distances (e.g., Manhattan or Mahalanobis), correlation-based distances (e.g., $D_{ij}=1-\mathrm{Cor}(g_i, g_j)$), or information-theoretic metrics may also be used. The criteria for merging clusters is known as the linkage. Simply put, for any two clusters, we wish to consider merging, we examine the pairwise distances $D_{ij}$ for the genes in the merged clusters; the linkage can be set to be the average, minimum ("single" linkage), or maximum ("complete" linkage) of the pairwise differences within the resulting cluster. The choice of which clusters to merge is then based on which cluster pairs yield the smallest linkage. At each iteration, those pairs of clusters are merged, forming a binary tree, and the number of clusters is determined by the height at which user cuts the tree.

However, while hierarchical clustering has a long history in microarray analysis, it is extremely sensitive to the choice of distance metric and the linkage method used to merge the clusters, since the "greedy" agglomeration causes slight inaccuracies to snowball. Hierarchical clustering should therefore be considered an exploratory tool rather than an analytical one.

### $k$-Means Clustering

The well-established $k$-means clustering [24] technique provides a more stable partition of the genes. The algorithm iteratively finds points that define the centers of globular clusters: starting with a user-specified number of clusters $k$, it selects $k$ genes at random as starting centroids, and clusters all the genes based on the centroid to which they are closest. For each of the resulting $k$ clusters, new centers are computed based on the mean expression of the genes assigned to that cluster. The genes are reclustered with respect to the new centroids, and the process is repeated until the clustering assignments converge. In addition to being much less error prone than hierarchical clustering, $k$-means is also considerably faster. As with hierarchical clustering, however, the user must specify the number of clusters (which in the case of genes means guessing at the number of "modules"). In addition, $k$-means performs poorly when the genes do not form globular, linearly separable clusters.

### Improved Approaches

To address these drawbacks, several refinements have been proposed. Graph-theoretic spectral techniques [25–30] are able to articulate clusters with nonlinear and nonconvex boundaries, allowing complex relationships between genes (such as those that oscillate differentially over the cell cycle) to be discerned. Variational clustering schemes [31, 32] achieve similar goals. Several schemes have also been

proposed to estimate the number of clusters from the data itself rather than relying on user input [30, 33–36]. Combined, these methods produce robust partitions even in complex data sets.

One interesting and extensible approach, consensus clustering [36], is a method that may be wrapped around any clustering algorithm of choice (hierarchical, $k$-means, spectral clustering, etc.) to provide both an estimate of the number of clusters present in the data and a measure of the robustness of the clustering. In consensus clustering, the data is randomly subset so that only a portion of the genes and samples are used. The clustering algorithm of choice is then used to cluster the samples or genes into $k = 2, 3, 4, \ldots$ groups for multiple random subsets of the data. For each $k$, a consensus matrix is obtained where the $i, j$th entries are the fraction of times gene $i$ and gene $j$ were assigned to the same cluster across multiple subsets. For a truly robust partition, it is expected that the entries will all be close to 1 or 0, that is, either $i$ and $j$ are consistently placed in the same cluster, or they are consistently placed in different groups. (This reflects the intuition that if the algorithm only place objects in the same cluster together half the time, it is questionable whether a separate cluster truly exists.) The optimal $k$ (number of clusters) is that for which the consensus matrix comes closest to the ideal of pure 1's and 0's. Wrapping consensus clustering around $k$-means or hierarchical clustering mitigates the limitations of those methods; moreover, because the consensus technique may be wrapped around any clustering engine, it can readily incorporate the advantages offered by the more sophisticated nonlinear clustering algorithms described above. Recently, consensus clustering was applied to identify molecular subtypes of diffused large B-cell lymphoma, leading to the identification of highly reproducible transcriptional signatures corresponding to differential signaling cascades [37].

## *Dimension Reduction*

As the number of genes assayed is vast, it is often of interest to find a small number of representative patterns that describe most of the variation observed in the data and on which the gene expression may be modeled, rather than dealing with the whole data set. This problem is closely related to clustering: by identifying dominant patterns of gene expression (across samples or over time), one may then find clusters of genes that match particular patterns. Those pattern-based clusters may then be examined for common regulatory elements.

### Principal Component Analysis

The simplest and best-known dimension reduction technique is principal component analysis (PCA) [38], which transforms a set of observations of possibly correlated variables (e.g., gene expression measurements) into a new set of variables, called the principal components (PCs), which are constructed such that the PCs are completely

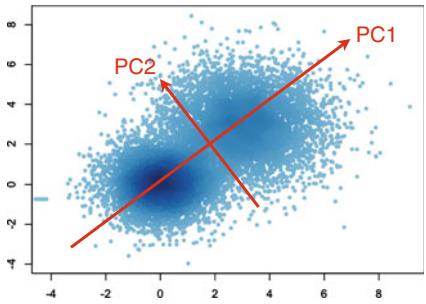


**Fig. 8.2** In principal components analysis, the principal components are defined such that the first principal component (PC1) lies along the direction of greatest variation and each succeeding component (in two dimensions, only PC2) is defined to lie in an orthogonal direction with the highest variance. Geometrically, the PCA space is a rotation of the original axes

independent of each other. The transformation is defined such that the first principal component lies along the direction of greatest variation in the data, accounting for as much of the overall variation in the gene expression between samples as possible. Each succeeding component lies, in turn, along the direction of the highest variance under the constraint that it will be orthogonal to (i.e., uncorrelated with) the preceding components. Mathematically, the principal components are the eigenvectors of the covariance matrix; the associated eigenvalues indicate the amount of variance along each component. A graphical illustration in two dimensions is given in Fig. 8.2.

The transformation is linear, that is, the original coordinates (genes) are rotated in the PCA space, such that the bulk of the variation lies along the first PC, and so on, as shown in Fig. 8.2. Each gene may thus be described using a weighted combination of components (and vice versa). Because the bulk of the statistical variation in the data is contained in the first few components, it is possible to use just the first few PCs, rather than the full $10^5$-dimensional feature space, when analyzing the data. The resulting clusters may then be examined for common regulatory elements. Recently, Chilarska and coworkers used this approach to identify combinatorial transcriptional control in a genome-wide study of blood stem/progenitor cells [39]. PCA has also been used to identify "fingerprints" of hematopoietic stem cell differentiation [40].

### Eigengenes

The principal components transformation can be written in terms of another matrix factorization called the singular value decomposition (indeed, computation of the principal components is typically done from the SVD, rather than the mathematically equivalent but computationally costly eigendecomposition of the covariance matrix). While PCA yields a matrix containing the PCs (i.e., the eigenvectors) and a vector of loadings (the eigenvalues), SVD yields *two* matrices, each describing an orthogonal basis, and a vector of so-called singular values. When applied to gene expression data, the two matrices have the dimensions of the genes and samples,

which are referred to as the "eigengenes" and "eigenarrays," respectively [41]. Like the principal components, the eigengenes (eigenarrays) are unique orthonormal superpositions of the genes (samples). Eigengenes/arrays that are inferred to represent noise may be filtered out, much like filtering out the higher PCs in PCA. Representing the data by the remaining eigengenes and eigenarrays gives a global picture of the dynamics of gene expression, in which individual genes (or samples) are clustered into groups of similar regulation and function (or similar cellular state and biological phenotype, respectively). These clusters may then be associated with observed genome-wide effects of regulators. Recently, this method has been used to uncover the combinatorial role of transcription factors regulating the yeast sulfur assimilation pathway [42] and combined with dynamical modeling; a similar approach could be used to link high-throughput data to dynamical models of blood stem cell fate (e.g., [43]).

**Nonlinear Dimension Reduction**

The patterns described by the principal component vectors or eigengenes are linear combinations of the gene expression measurements. However, if the biological patterns of interest have a nonlinear form, as is likely to arise from regulatory networks with feedback loops, neither classical PCA nor SVD can articulate those patterns. Instead, nonlinear dimension reduction (NLDR) techniques must be used. NLDR may be thought of as a nonlinear version of PCA where the coordinates are "threaded" along the direction of greatest variability. Optimally detecting those paths is a mathematically and computationally challenging task, and several methods have been proposed including kernel PCA, Laplacian eigenmaps, IsoMaps, and spectral embedding [44, 45]. Of these approaches, the neural-network-based self organizing map (SOM) [46] is the best represented in the genomics literature. Figure 8.3 provides an illustration of the first SOM coordinate versus the first PC for data lying on a curved manifold; while the first PC captures only 76.77 % of the variance, the first component of the SOM captures 93.14 % and provides a better description of the underlying pattern.

This property makes SOM (and NLDR generally) particularly well-suited for analyzing transcription dynamics, where the relationships between genes may not be strictly linear. SOM has been applied to detect and interpret gene expression patterns governing hematopoiesis [47]. For an in-depth mathematical treatment of various NLDR methods, the reader is referred to [48].

However, while NLDR provides a more accurate and possibly more biologically meaningful dimension reduction than PCA or SVD, it must also be noted that the transformation from the new, dimension-reduced space to that of the genes is not a straight-forward (or even necessarily possible) task. This is a direct consequence of their nonlinearity and places them in stark contrast to PCA and SVD, from which it is easy to recover the original coordinates. This, in turn, means that it is very difficult to say which genes are the ones driving the dominant pattern observed, which can pose problems when it comes time to identify specific genes for validation work. In short,

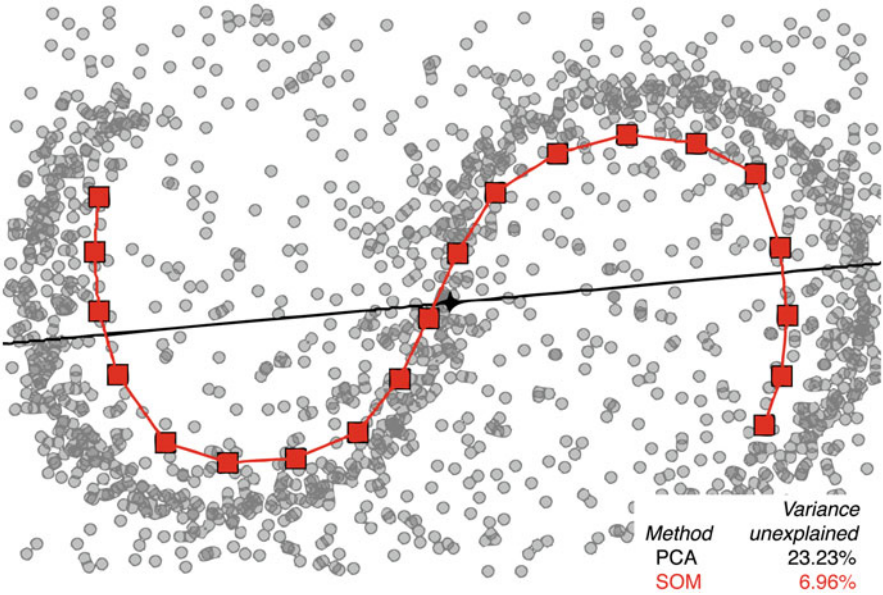


**Fig. 8.3** Comparison of SOM versus PCA. While the first PC captures only 76.77 % of the variance, the first component of the SOM captures 93.14 % and provides a better description of the underlying pattern

what we gain in accuracy and representativeness in NLDR is lost in interpretability. The choice of dimension reduction should thus be undertaken with the end goal of the analysis in mind.

## *Pathway Analysis*

Pathways, or networks of functionally related genes and molecules, provide a natural framework in which systems-level effects may be investigated in the context of existing "expert" knowledge. Pathway definitions may be extracted from a growing number of databases, including the Pathway Interaction Database [49], KEGG [50], Reactome [51], and InnateDB [52], among others. Many statistical computing packages, including R/BioConductor, have interfaces to these databases [53].

Analyzing high-throughput molecular measurements at the pathway level have two significant benefits. First, it permits the grouping of hundreds of thousands of genes (or other biomarkers) into several hundred pathways, reducing the complexity of the analysis. Second, identifying active pathways that differ between two conditions can provide more explanatory power and mechanistic insights than a simple list of genes. These benefits have given rise to a vast number of different pathway analysis approaches over the past decade [54].

Many tools for pathway analysis are available, including free, open-source R software from the BioConductor project [13] (http://www.bioconductor.org) and popular commercial tools such as Ariadne Genomics Pathway Studio (http://www.ariadnegenomics.com/) and Ingenuity Pathway Analysis (http://www.ingenuity.com/). As these tools are well-documented and constantly evolving, we focus here on the underlying methodology.

### Enrichment Analyses

The simplest pathway analysis approach is an overrepresentation analysis, which seeks to address statistically the following question: given a set of genes known to be on a pathway, and given the list of genes detected to be different in the study (e.g., with FDR$\leq$0.05 in a test of differential expression), is there greater overlap than would be expected by chance alone? That is, do the significant genes appear to aggregate in certain pathways? The probability of having an overlap of $m$ or more genes when there are $M$ significant genes out of $N$ genes assayed, and $n$ genes in total on the pathway is given by the hypergeometric distribution,

$$\Pr(X \geq m|N, M, n) = \sum_{r=m}^{n} \frac{\binom{M}{r}\binom{N-M}{n-r}}{\binom{N}{n}} \tag{8.1}$$

which is easily computed for all gene sets of interest.

While simple, overrepresentation analysis has a significant limitation: because it uses only the most significant genes (e.g., those passing the arbitrary FDR$\leq$0.05 threshold), marginally less significant genes (e.g., FDR = 0.051) are discarded, resulting in information loss. In contrast, the popular Gene Set Enrichment Analysis (GSEA) algorithm [55, 56] uses the full list of all genes, ranked in order of significance, and uses a Kolmogorov–Smirnov running sum statistic to answer the question: what is the probability that the genes in this pathway lie as near the top of the ranked list as we observe them to be? Significance may be computed either by permuting the sample labels or permuting the genes included in the pathway [54, 57, 58]. These methods have been applied successfully to a variety of studies, including expression profiling of acute lymphocytic leukemia subtypes [59]; pathways involved in the activation of memory T cells, monocytes, and B cells [60]; and resistance to chemotherapy in acute myeloid leukemia cells [2].

Nevertheless, both simple overrepresentation analysis and GSEA have a common drawback: they rely upon the computation of gene-level statistics. It is well known that complex diseases exhibit considerable molecular heterogeneity, either due to causative mechanisms that can be deleteriously affected in a variety of ways (such that no particular alteration is dominant among the case samples) or to those that are only deleteriously affected through a specific combination of particular alterations (such that control samples may have some, but not all, the alterations necessary to produce the case phenotype). As a result, individual genes may fail to reach significance in univariate tests of significance, and pathway analyses that rely on

single-gene association statistics may fail to detect significant pathways simply because the constituent genes do not exhibit *univariate* associations. Simply put, by relying on univariate, gene-level tests, overrepresentation and enrichment analyses still make the reductionist assumption that regulatory systems may be investigated by examining their parts in isolation.

### Pathway Summary Statistics

An alternative is to compute a pathway-level "summary" statistic: a single value that summarizes the expression level (or other data) for all the genes in the pathway. For each pathway, a summary statistic is obtained for each sample based on its profile, and those summaries may then be tested for association with the condition of interest. A very crude example is to use the average expression value for the genes on the pathway, such that a sample in which many of the genes are upregulated will have a high pathway summary value, regardless of which genes happen to be upregulated. However, simply averaging the gene expression levels is a poor measure of pathway activity from a biological point of view, since these mechanisms involve both up- and downregulation for which coordinated gene expression (and hence correlations) are important. A more justifiable approach, therefore, is to use PCA for the genes in the pathway of interest, selecting the first PC as the "pathway summary." This has the effect of summarizing the expression (or other) values for all the genes on the pathway by a single number that describes the bulk of the variation and which mathematically accounts for the correlations in gene expression. This technique has been used identify differential pathways in leukemia [61] as well as other cancers.

Extending this idea, we proposed a method [62] in which the pathway summary values and genes not known to be on the pathway were tested for differential correlation. In the "Gene × Pathway Correlation (GPC) Score" method [62], we first computed pathway summary values based on the first PC for every pathway of interest, yielding for each pathway $j$ a value $p_{j,m}$ summarizing sample $m$'s expression across pathway $j$'s genes. For each gene $i$ with expression $g_{i,m}$ in sample $m$, we compute the GPC-score as the difference in the correlations of $g$ and $p$ in the case and control phenotype,

$$\text{GPC-score} = \text{Cor}_{m\in\text{Case}}(p_{j,m}, g_{i,m}) - \text{Cor}_{m\in\text{Control}}(p_{j,m}, g_{i,m}), \tag{8.2}$$

yielding a gene × pathway matrix of correlation differences for each gene-pathway pair. The significance of the correlation differences are assessed by randomly permuting the case and control labels. This method has the power to identify new regulatory interactions (by finding correlated gene-pathway pairs), as well as to detect those which are altered in disease (by identifying gene-pathway pairs with significant differences). An example gene-pathway pair identified in a prostate cancer study is given in Fig. 8.4.

Although this method was applied in [62] exclusively to mRNA expression data, the same technique may be applied as an integrative analysis using both mRNA

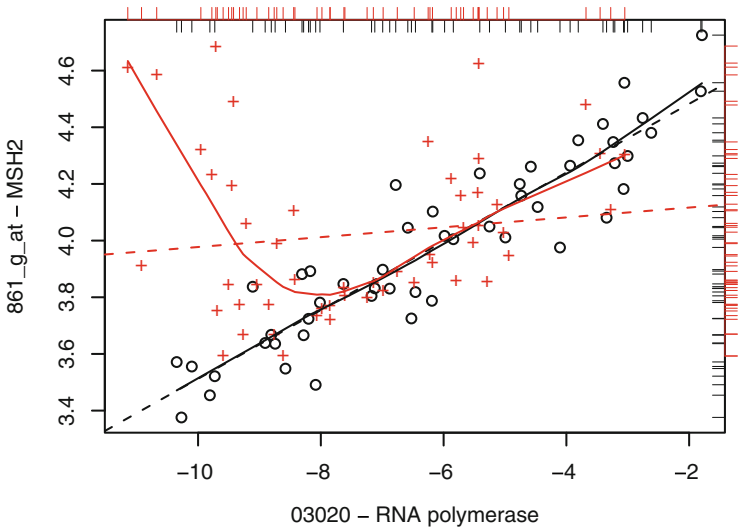


**Fig. 8.4** GPC-Score identifies differential gene-pathway coexpression for the MSH2 (mismatch repair) gene and the RNA polymerase pathway for a subset of prostate tumor samples; these samples corresponded to worse clinical outcomes. (Image: [62])

and other genomic or environmental measurements. For instance, one can apply it to combined miRNA/mRNA data to search for differentially correlated miRNA-pathway pairs, thereby identifying miRNAs whose expression modulates the activity of regulatory networks.

These pathway summaries effectively amount to selecting the $n$ genes on a given pathway and applying a dramatic dimension reduction to go from the $n$ features down to a single one. As such, the same caveats about linearity described in Section "Dimension Reduction" apply, namely, linear methods cannot account for complex or oscillatory relationships between genes. Instead, NLDR such as kernel PCA or Laplacian eigenmaps may be used, providing a more accurate and biologically representative summary of the expression patterns across a pathway.

## Sample Class Prediction and Class Discovery

In the previous section, our goal was to categorize genes into biologically relevant functional modules, either by grouping the genes into clusters of correlated expression or by pathway analysis. Here, we turn our attention to categorizing *samples* based on complex patterns in the experimental data, with the goal of predicting the status or outcome for a new sample or discovering sample subclasses that were previously unknown.

The gene-level tests described in Section "Generating High-Throughput Data" yield lists of differentially expressed genes and significantly associated genetic variants that are ubiquitously reported in genomic studies. However, while these genes are significantly associated with the phenotype of interest, they may not accurately classify or predict the outcome for a new sample. To develop predictors from high-throughput data, machine-learning algorithms are commonly used. These algorithms are first "trained" against a subset of the data for which the outcomes are known, and then evaluated for accuracy against an independent "testing" subset (for which the outcomes are also known). If the classifier performs well in the testing subset, it may then be validated in a distinct data set. The procedure of dividing the data into training/testing sets, known as cross-validation, ensures that the models are not overfit to technical nuances in the data. As the known sample labels (case/control, exposed/unexposed, etc.) are used to train the machine, these techniques are referred to as "supervised" classifiers.

The literature now contains many supervised machine learning algorithms; for a deep and comprehensive exposition, the reader is directed to Hastie and Tibshirani's *Elements of Statistical Learning* [48]. Here, we discuss three powerful techniques: one designed for continuous data (such as gene expression), one designed for categorical (SNP or sequence) data, and a third that can accommodate a mixture of covariates. From a systems-biology perspective, the predictive "signatures" obtained from these algorithms may be used to suggest functional modules, identify epistatic interactions between genetic variants, or provide an integrative analysis that combines genomic, epigenetic, and expression data. We also discuss unsupervised methods for class discovery (i.e., the identification of sample subtypes based solely on the high-throughput data). Such methods can articulate complex, systems-level similarities and differences that would be indetectable by association tests alone.

## *Nearest Shrunken Centroids*

Given a set of samples comprising different categories (cases and controls, say), and a new sample whose categorization is unknown, a natural way to classify it is to ask which group, on average, the unknown sample is closer to. This approach is referred to as a "nearest centroid classifier"—the centroids represent the average gene expressions (or other data) in each sample class, and the new sample is classified based on the centroid to which it is nearest.

In the context of genome-wide expression profiling, the centroid for each sample class resides in a very high-dimensional space. If $10^5$ genes have been assayed, the centroid for the cases is a $10^5$-dimensional vector which gives the average expression for each gene across all the case samples; likewise the controls. As the vast majority of these genes are not biologically relevant, it is important to remove those which can be reasonably assumed to be noise. One powerful approach is to consider not only the average gene expression, but the variance as well, moving the centroid coordinates closer to each other by an amount proportional to the variance of the corresponding genes [63]. This procedure is referred to as "shrinking" the centroids.

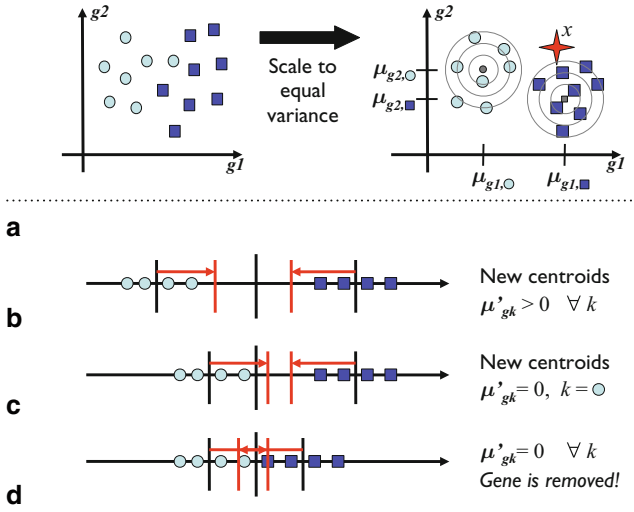


**Fig. 8.5** Nearest shrunken centroids classifier. In **a**, the nearest centroid classifier in two dimensions is illustrated. There are two classes of samples $k$, shown as *light circles* and *dark squares*. After scaling each gene (here, $g1$ and $g2$) to unit variance within each group $k$, the unknown sample $x$ is classified based upon the nearest centroid $\mu$ (in this case, the dark squares). (b)–(d) illustrate the shrinkage of the centroids for a gene $g$. Centroids $\mu_{gk}$, shown as a black line, are moved in the direction of the center line to a new position $\mu'_{gk}$. In **b**, neither cross the center line, and the new position is retained. In **c**, the centroid for the light circles crosses the center line and is thresholded to 0. In **d**, both centroids cross the center line and are thresholded to 0; because the new centroids are equal, the gene no longer contributes to the classification

A graphical illustration is given in Fig. 8.5. Mathematically, the nearest centroids procedure attempts to classify a new sample with gene expression $\mathbf{x}$ into group $k$ such that $\delta_k(\mathbf{x})$ is minimized:

$$\boldsymbol{\delta}_k(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \ln(\boldsymbol{\pi}_k), \tag{8.3}$$

where $\boldsymbol{\mu}_k$ are the centroids for each group $k$, $\boldsymbol{\Sigma}$ is the covariance matrix (across all groups), and $\boldsymbol{\pi}_k$ are the prior probabilities that $\mathbf{x}$ belongs to each group $k$. For instance, when building a classifier to detect a rare disease, $\boldsymbol{\pi}_k$ may be taken from the disease prevalence in the population, reflecting the low probability that the patient has the disease of interest.

In the "shrunken" approach [63], the $g$th entry of the vector $\boldsymbol{\mu}_k$ (i.e., the mean of gene $g$ in group $k$) is moved from $\mu_{gk}$ to $\mu'_{gk}$ by an amount proportional to the pooled variance $s_g$ (plus a slight offset $s_0$) for gene $g$:

$$\mu'_{gk} = \mu_{gk} - \Delta(s_g + s_0)\sqrt{1/n_k + 1/n}, \tag{8.4}$$

where $n_k$ is the number of samples in group $k$, $n$ is the total number of samples, and the degree of shrinkage is controlled by the parameter $\Delta$. Genes that cross

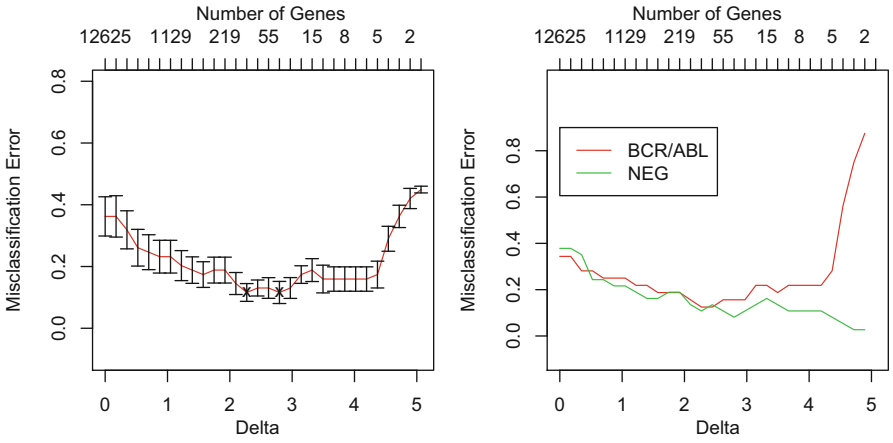


**Fig. 8.6** Application of the nearest shrunken centroids classifier to distinguish cytogenetically normal cases ("NEG") from those with BCR/ABL fusion based on gene expression profiles of patients with acute lymphoblastic leukemia (ALL). The overall misclassification error is shown on the left, while the misclassification error for the known groups is shown on the right. As the shrinkage parameter $\Delta$ increases, fewer genes remain in the model. Initially, the removal of genes improves the accuracy as "noisy" genes are removed. Optimal values of $\Delta$, corresponding to the smallest error observed in the cross-validation, are obtained at $\Delta = 2.272$ (115 genes) and $\Delta = 2.796$ (40 genes). Increasing $\Delta$ beyond 3 removes informative genes (only 20 remain at $\Delta = 3$), causing a dramatic increase in the error rate, particularly amongst BCR/ABL cases

the "overall" centroid across all groups (i.e., those for which $\mu'_{gk} - \mu_g \leq 0$) for all classes $k$ do not contribute to the final model, resulting the removal of high-variance "noisy" genes from the classifier. The shrinkage parameter $\Delta$ controls the aggressiveness of the removal (higher $\Delta$ forces a greater degree of shrinkage and hence more genes are removed), and is optimized by cross-validation. The data set is split into multiple training and testing subsets, and samples in the testing subset are classified according to the shrunken centroids in the training subset. By varying the value of $\Delta$ over multiple training/testing splits, it is possible to choose $\Delta$ such that the error in the testing subset is minimized.

An example applied to gene expression data from well-known study [64] of acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) is given in Fig. 8.6. Here, the nearest shrunken centroids classifier, implemented in the R package `pamr` [65], has been applied to data from 12,625 genes in 95 ALL cases, of whom 42 are cytogenetically normal and 37 have BCR/ABL fusion. The classifier was trained using tenfold cross validation with increasing values of $\Delta$ ranging from $\Delta = 0$ (no shrinkage, all genes used) until no genes remained at $\Delta \approx 5$. As shown in Fig. 8.6, the misclassification error initially drops as $\Delta$ is increased and noisy genes are removed. The optimum $\Delta = 2.796$ yields an error rate of approximately 13 % using 40 genes. Further increasing $\Delta$ has the effect of removing informative genes, causing the error rate to rise again. This is particularly true for the BCR/ABL cases, which are frequently misclassified as cytogenetically normal as $\Delta$ is increased above the optimum.

## *Identifying Epistatic Interactions*

As with expression profiling by microarray and NGS, genome-wide association studies (GWAS) have become a powerful and increasingly affordable tool to study genetic sequence variants associated with disease. Modern GWAS yield information on millions of single nucleotide polymorphism (SNPs) loci distributed across the human genome, and have already yielded insights into the genetic basis of complex diseases [66, 67]; a complete list of published GWAS can be found at the National Cancer Institute-National Human Genome Research Institute (NCI-NHGRI) catalog of published genome-wide association studies [68]. As described above, the data is typically analyzed by testing the alleles at each locus for association with case status; significant association is indicative of a nearby genetic variant which may play a role in the phenotype being studied. Genomic regions of interest may also be investigated by haplotype analysis, in which a handful of alleles transmitted together on the same chromosome are tested for association with disease; in this case, the loci which are jointly considered are located within a small genomic region, often confined to the neighborhood of a single gene.

Recently, however, there has been increasing interest in multilocus, systems-based analyses. This interest is motivated by a variety of factors. First, few loci identified in GWAS have large effect sizes (the problem of "missing heritability") and it is likely that the common-disease, common-variant hypothesis [69, 70] does not hold in the case of complex diseases. Second, single marker associations identified in GWAS often fail to replicate. This phenomenon has been attributed to underlying epistasis [71], and a similar problem in gene expression profiling has been mitigated through the use of gene-set statistics. Most importantly, it is now well understood that because biological systems are driven by complex biomolecular interactions, multigene effects will play an important role in mapping genotypes to phenotypes; recent reviews by Moore and coworkers describe this issue well [70, 72]. In addition, the finding that epistasis and pleiotropy appear to be inherent properties of biomolecular networks [73] rather than isolated occurrences motivates the need for systems-level understanding of human genetics.

Several multi-SNP GWAS analysis approaches have been described in the literature. Thorough reviews are provided in [74, 75], and we briefly describe several here. Building on the well-established Gene Set Enrichment Analysis [55] method initially developed for gene expression data, two articles have proposed an extension of GSEA for SNP data [76, 77] using the $\chi^2$ SNP-level statistics. As in expression-based GSEA, the reliance on single-marker statistics means that systematic yet subtle changes in a gene set will be missed if the individual genes do not have a strong marginal association. In the case of a purely epistatic interaction between two SNPs in a set, the set may fail to reach significance altogether.

As an alternative, the notion that cases will more closely resemble other cases than they will controls has motivated a number of distance-based approaches for detecting epistasis. Multi-dimensionality reduction (MDR) has been proposed and applied to SNP data [78–80]. The technique is conceptually similar to the nearest Shrunken centroids classifier described above; here, sets of $l$ SNPs are exhaustively searched

for combinations that will best partition the samples by examining the $3^l$ cells in that space (corresponding to homozygous minor, heterozygous, and homozygous major alleles for each locus) for overrepresentation of cases. While this method finds epistatic interactions without requiring marginal effects and can be structured to incorporate expert knowledge, it is limited by the fact the the total number of loci to be combinatorially explored must be restricted to limit computational cost. To address this, an "interleaving" approach in which models are constructed hierarchically has been suggested [79] to reduce the combinatorial search space. A recent and powerful MDR implementation [81] taking advantage of the CUDA parallel computing architecture for graphics processors has made feasible a genome-wide analysis of pairwise SNP interactions. Still, MDR remains computationally challenging, such that expanding the search to other SNP set sizes (rather than restricting to pairwise interactions) can be impeded by combinatorial complexity if an exhaustive search is to be performed.

In order to narrow down the combinatorial complexity of discovering SNP sets using techniques such as MDR, feature selection may be employed. Of particular importance here is the distance-based approach of the Relief family of algorithms [82–85]. These are designed to identify features of interest by weighting each feature through a nearest-neighbor approach. The weights are constructed in the following way: for each SNP, one selects samples at random and asks whether the nearest neighbor (across *all* SNPs) from the same class and the nearest neighbor from the other class have the same or different values from the randomly chosen sample. Attributes for which in-class nearest neighbors tend to have the same value are weighted more strongly as being more representative of the underlying biology. Because the neighbor distances are computed across all attributes, Relief-type algorithms can identify SNPs that form part of an epistatic group and provide a means of filtering out unpredictive loci.

While these methods have so far been applied to finding small groups of interacting SNPs, one may instead be interested in whether cases and controls exhibit differential distance when considering a large number of genes. A multi-SNP statistic has been proposed in the literature [86–88] for determining whether a new sample of interest is on average (across a large number of SNPs) "closer" to one population (e.g., cases) than to another (e.g., controls). The method [86] is motivated by the idea that a subtle but systematic variation across a large number of SNPs can produce a discernible difference in the closeness of an individual to one population sample relative to another. Assuming an individual $Y$ and two population groups $F$ and $G$ with minor allele frequencies $y_i$, $f_i$, and $g_i$ ($y_i \in \{0, 0.5, 1\}$) for SNP $i$, respectively, we write the distance metric for SNP $i$ as

$$D_{Y,i} = |y_i - f_i| - |y_i - g_i|\,, \tag{8.5}$$

and then consider the normalized mean across all SNPs of interest:

$$S_Y = \frac{\mathrm{E}D_{Y,i}}{\sqrt{\mathrm{Var}D_{Y,i}/l}}\,. \tag{8.6}$$

An illustration is given in Fig. 8.7.

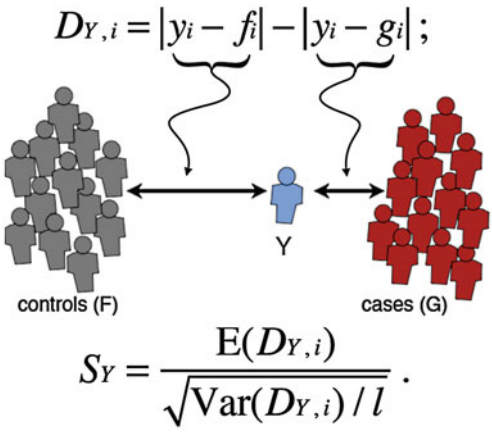


**Fig. 8.7** Genetic distance metric (Eqs. 8.5–8.6). If $Y$ is closer to $G$ than to $F$ for locus $i$, $D_{Y,i}$ is positive. If $D_{Y,i}$ is consistently positive across all $l$ loci, $S_Y$ will be so as well, indicating a tendency for $Y$ to have more "$G$-like" patterns of genetic variation

It is clear from Eqs. 8.5 and 8.6 that individuals $Y$ whose minor allele frequencies at locus $i$ more closely resemble those of group $G$ will have a positive $D_{Y,i}$ and vice versa. By chance alone, we would expect $D_{Y,i}$ to be as frequently positive as negative, yielding $S_Y \approx 0$. However, a slight but consistent tendency to be closer to one group than another across a set of SNPs will cause deviations in $S_Y$ (Fig. 8.7). The significance of $S_Y$ in Eq. 8.6 may be assessed either parametrically by assuming normality (only in the case of large $l$), or by resampling the $F$ and $G$ populations.

While this statistic was originally designed to identify group membership of individuals who were known to be in either $F$ or $G$ (and hence contributing to $f_i$ and $g_i$), it was later shown in [87] that even out-of-pool breast cancer cases were in general "closer" to the population of other cases than to the controls, suggesting that the combination of multiple alleles has the potential to classify *new* samples. Building on these ideas, the PoDA [89] technique has been proposed to find pathway-based SNP-sets that distinguish cases from controls. The hypothesis is that if the SNPs come from a pathway that plays a role in disease, there will be greater in-class similarity than between-class similarity in the genotypes for those SNPs, i.e., a case will show greater genetic similarity to other cases than to controls for the SNPs on a disease-related pathway, but will be equidistant for the SNPs on a non-disease-related pathway. In order to identify the significant pathways, a leave-one-out cross-validation procedure is used: each sample in the study is treated as unknown, and the pathways with SNPs that most accurately classify the "unknown" samples are flagged. Because subtle but consistent $D_{Y,i}$'s will accumulate to give large values of $S_Y$ in Eq. 8.6, PoDA can identify multi-SNP sets which differ systematically even when the single-SNP associations are not strong enough to be significant, making it useful for detecting epistatic interactions. By restricting the PoDA SNP sets to those defined based on known relationships (e.g., SNPs in genes sharing a common pathway), one may incorporate expert knowledge to reduce the search space and provide biological interpretability.

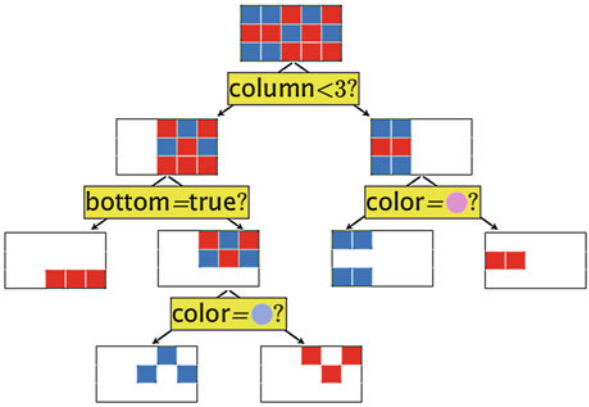


**Fig. 8.8** Schematic of a decision tree. At each step, a variable and threshold is chosen to optimally partition the samples based on known labels. The decision rules may operate on continuous variables (like color here, with *blue* and *red* coming closest to the *mauve* and *periwinkle* ideals, respectively), categorical variables (like column, which can take on values 1–5), or booleans (like "bottom," which is either true or false). The partitioning stops when the nodes are pure. Variables may be used multiple places in the tree (such as color here), so long as they are not used along the same branch twice

## *Random Forests*

The methods described above are designed to be applied to one type of data from a single experimental modality—continuous data, such as that obtained in expression profiling, or categorical data, such as that obtained by SNP or sequencing studies. Now that genome-wide experiments are growing increasingly affordable, it is becoming more common for a variety of assays to be run on the same set of samples, allowing the various measurements to be integrated in the analysis. The Random Forests algorithm [90] is a decision-tree based classifier that permits multiple types of data to be mixed a priori, enabling its use as an integrative predictor.

Decision trees are a conceptually simple supervised classifier. At each step, a variable and threshold is chosen to optimally partition the samples based on known labels. The decision rules may operate on continuous or categorical variables. The partitioning continues until either all nodes are pure or all variables have been used. An illustration is given in Fig. 8.8. Once the rules are established based on labeled samples (i.e., the tree is trained), the rules may be used to classify a new sample of unknown status.

Because at every level the decision tree partitions the samples completely, certain partitions are not possible to achieve. An example is given in Fig. 8.9; here, there is no way to place the cuts (corresponding to decision rules) to isolate the sample in the center and achieve pure partitions.

In order to address this issue, the "Random Forests" classifier was proposed [90]. As in consensus clustering (described above), in Random Forests, we also randomly

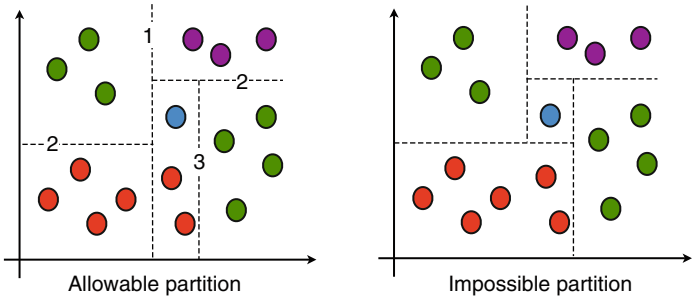


**Fig. 8.9** Possible and impossible decision tree partitions. On the left, a possible partition (at levels 1, 2, and 3 in the decision tree) is shown; on the right, a partition that cannot be achieved with the classical decision tree algorithm

subset the data, selecting samples on which to grow the tree using a random sample of features (gene expression, SNP alleles, clinical covariates, or any other available information). The procedure is repeated for many different samplings, yielding a "forest" of decision trees, each trained on a random subset of the data (in much the same way that one obtains a multitude of clusterings of randomly subsetted data in consensus clustering). New samples are then classified according to a majority vote of the trees.

As the goal here is to *predict* rather than cluster, the measure of accuracy is not how well-clustered the selected samples are (as it is in consensus clustering), but how well the decision tree predicts the status of the samples *not* selected in the random sample. For each tree, one can compute the the prediction accuracy for the "out of bag" (OOB) samples—those not used in the training of that particular tree. The average OOB error rate is considered to be a good estimate of the testing error, since each OOB error rate computation is based on samples not used in that particular tree. The OOB error rate is also used to tune the size of the random subset of features. The more features are kept, the more similar the trees will be to one another (eventually converging to identical trees), leading to a forest that may be overfit. The smaller the size of the feature subset, the more diverse the trees are, but each tree will exhibit worse per-tree performance due to the lost information. By varying the size of the feature subset and examining the OOB error rate, these two competing forces may be optimally balanced to yield a forest of trees that are neither underdetermined nor overfit.

Random forests have a number of useful features as an integrative predictor: it can incorporate different data modalities, is invariant to transformations of the data, can handle missing data easily, has a tuning mechanism to prevent overfitting, and provides an estimate of its accuracy. In addition, by looking at the purity of the partition each time a particular feature is used across the entire forest of trees, one may obtain a measure of its importance, yielding a ranked list of discriminatory markers. The ranked list may then be used as an input to pathway enrichment analyses (see Section "Pathway Analysis"), providing further systems-level insight [91]. This approach both allows one to combine data types in the pathway analysis and indicates

pathways that are not only "hit" by differential genes but by those that are *predictive* of the biological outcome. Alternatively, pathway summary statistics (as described in Section "Pathway Analysis") may be used as the features input to the Random Forests algorithm.

These features make Random Forests a powerful and highly accurate [92–94] algorithm for generating predictive models. Recently, it has been applied to public expression data as an in-silico screen to discover agents that eradicate leukemia stem cells [95]; applied to a SNP study to identify genomic variants that govern progression-free survival of myeloma patients [96]; and to elucidate transcription factor activity in hematopoietic stem cell differentiation [97].

## *Class Discovery*

The prediction algorithms described above rely upon supervised training using a set of samples for which the true classification is known. However, as with clustering, our knowledge about the true structure of the data may be incomplete in the sense that there exist subtypes which are either unknown or not reflected by the training sample labels. In particular, if a set of samples comprise a single clinical phenotype but span several different molecular subtypes, classifying a new sample based on molecular data may be highly error-prone owing to the lack of a distinct pattern in the training set. As a result, it is often of interest to attempt to discover any existing molecular subtypes present in the data. To this end, the clustering and cluster-number determination algorithms described in Section "Identifying Functional Modules" may be applied to samples (as well as to genes) to discover the optimal number of sample clusters. As with genes, it is important to recognize that these clusters may not be linearly separable, and therefore nonlinear techniques are likely to be more accurate [98]. The application of these techniques may then be followed by training a supervised classifier on the detected molecular subtypes. (Note that if a nonlinear clustering method is used, it is necessary to ensure the appropriate nonlinear out-of-sample extension is used to project the test samples onto the nonlinear space defined by the training samples, as discussed in [44].)

### The Partition Decoupling Method

One approach for identifying molecular subtypes at progressively finer scales without imposing linearity constraints is the partition decoupling method (PDM) [26, 30]. The PDM is able to reveal relationships between samples based on multigene expression profiles without requiring that the genes be differentially expressed (i.e., without requiring the samples to be linearly separable in the gene-expression space), as illustrated in Fig 8.10, and has the power to reveal relationships between samples at various scales, permitting the identification of phenotypic subtypes. The PDM consists of two iterated components: a spectral clustering step, in which the correlations

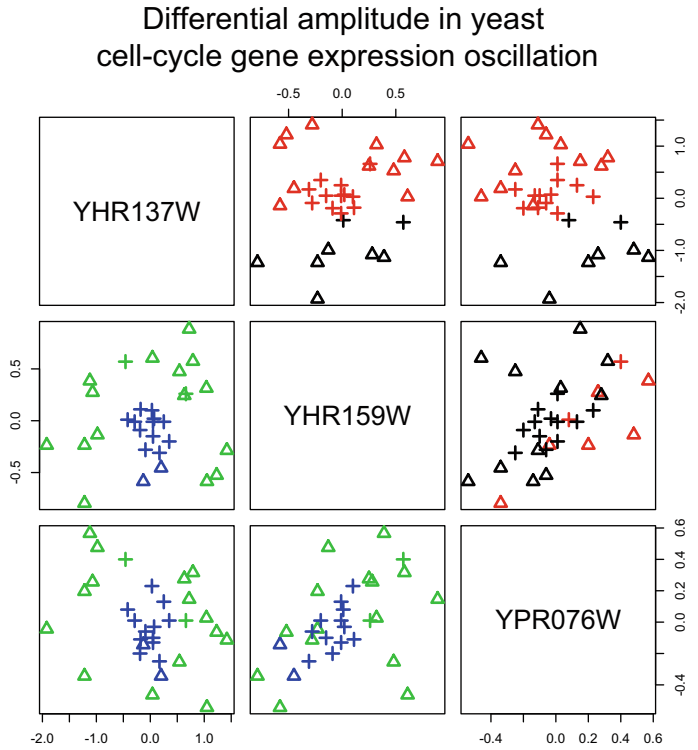


**Fig. 8.10** Expression levels for three oscillatory yeast cell-cycle genes from two different treatments: +, elutriation-synchronized samples; Δ, CDC-28 synchronized samples. The samples have different amplitudes of expression oscillation, leading to a "bullseye" pattern (note that the means for each gene in the two groups is approximately the same). Cluster assignment for each sample is shown by color for linear $k$ means clustering (*red/black*) above the diagonal, and nonlinear spectral clustering (*blue/green*) below the diagonal. Note the difference in accuracy. (Image: [30])

between samples in the high-dimensional feature space are used to partition samples into clusters, followed by a scrubbing step, in which the projection of the data onto the cluster centroids is subtracted so that the residuals may be clustered. As part of the spectral clustering procedure, a low-dimensional nonlinear embedding of the data is used, both reducing the effect of noisy features and permitting the partitioning of clusters with non-convex boundaries. The clustering and scrubbing steps are iterated until the residuals are indistinguishable from noise as determined by comparison to a resampled null model. This procedure yields "layers" of clusters that articulate relationships between samples at progressively finer scales.

The PDM has a number of satisfying features. The use of spectral clustering allows identification of clusters that are not necessarily separable by linear surfaces (such as the "bullseye" pattern in Fig. 8.10), permitting the identification of complex relationships between samples. This means that clusters of samples can be identified even in situations where the genes do not exhibit differential expression, a trait that

makes it particularly well-suited to examining gene expression profiles of complex diseases. The PDM employs a low-dimensional embedding of the feature space, reducing the effect of noise in the data. As the data itself is used to determine both the optimal number of clusters and the optimal dimensionality in which the feature space is represented, the PDM provides an entirely unsupervised method for class discovery, without relying upon heuristics. Importantly, the use of a resampled null model to determine the optimal dimensionality and number of clusters prevents clustering when the geometric structure of the data is indistinguishable from chance. By scrubbing the data and repeating the clustering on the residuals, the PDM permits the resolution of relationships between samples at various scales; this is a particularly useful feature in the context of gene-expression analysis, as it permits the discovery of distinct sample types, subtypes, etc.

These features make the PDM a powerful tool for genomic data analysis. As we demonstrated in [30] and illustrated here in Fig. 8.11, PDM detects with near-perfect accuracy both the phenotype and exposure groups in a study of radiation response; application to a leukemia data set with "incomplete" sample labeling demonstrated the PDM's ability to detect ALL subtypes simply from the expression data alone, with higher accuracy than other algorithms.

As described in [30], the accuracy of the PDM can be applied to gene subsets defined by pathways to identify mechanisms that permit the partitioning of phenotypes. In Pathway-PDM, one subsets the genes by pathway, applies the PDM, and tests whether the unsupervised PDM cluster assignments reflect the known sample classes. Pathways that permit accurate partitioning contain genes with expression patterns that distinguish the classes, and may be inferred to play a role in the underlying biology. As the PDM does not require the pathway's constituent genes be differentially expressed, complex regulatory relationships within pathways may be detected (such as those giving rise to the pattern seen in Fig. 8.10). It was further demonstrated in [30] that this approach, due to its increase accuracy, is a useful meta-analytical tool that can improve cross-study concordance, allowing more robust findings to be culled from existing high-throughput datasets.

## Network-Based Approaches

Further refinements to the analyses described here are achieved by examining the structure of interaction networks, rather than treating pathways as simple collections of genes. Network-based analyses fall into two broad categories: statistical analyses of high-throughput data in the context of putative interaction networks, and the inference of network topology from the data itself. A comprehensive introduction to network analysis in general may be found in [99]; case studies of its application to gene expression data are presented in [7], and a review of graph-theoretic concepts in the context of biology is provided in [100]. Here, we discuss a number of promising techniques.

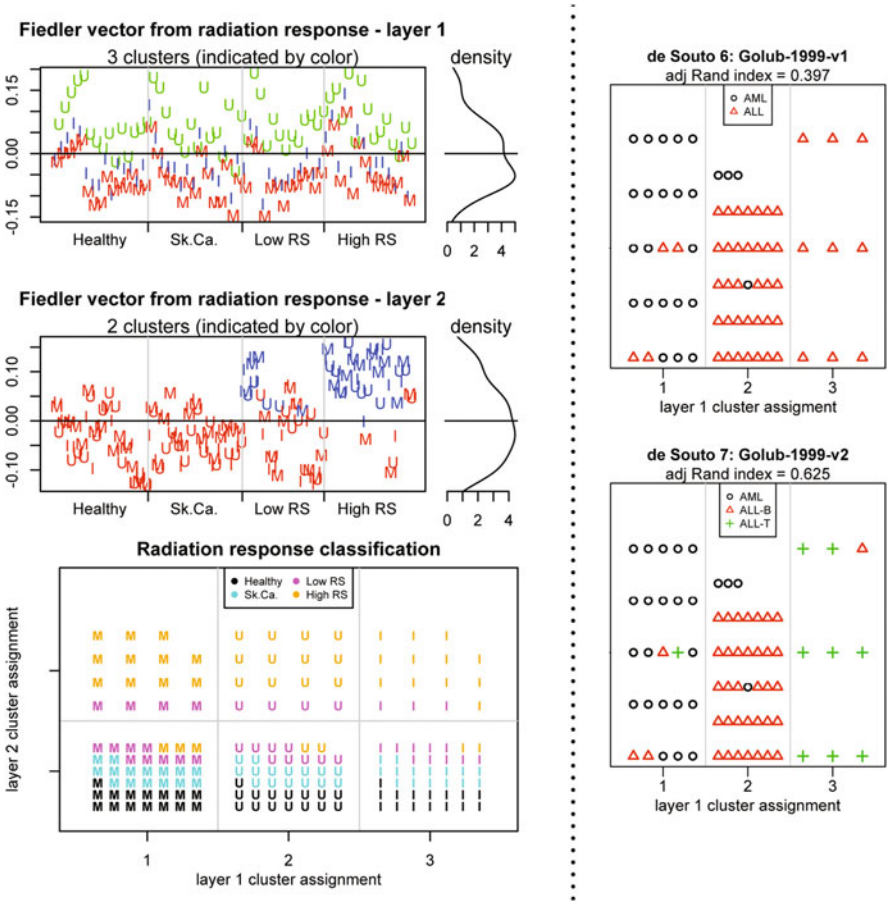


**Fig. 8.11** Multilayered, highly accurate unsupervised class discovery using PDM. Left, two "layers" of clusters correspond to the radiation exposure (*U*V light, *I*onizing radiation, *M*ock) and the case (high-RS) group (versus three control groups) in a radiation sensitivity study. The number of clusters in each layer is determined by the PDM itself from the data yielding three clusters in the first layer (*top left panel*) and two in the second (*center left panel*); the resulting classification is near-perfect discrimination of both phenotype and exposure (*bottom left panel*). Right, we see the clustering for leukemia data from [64]. The PDM automatically detects three clusters; in the top panel, comparison against the provided labels (*AML/ALL*) shows that the ALL group has been split by PDM; in the lower panel, it is revealed that this corresponds to a subtype difference (*ALL-B, ALL-T*), demonstrating PDM's ability to identify sample subtypes even when they may be unknown or unannotated in the data. (Image: [30])

## *Network Statistics*

To incorporate known interaction network topology with traditional pathway analyses (Section "Pathway Analysis"), several approaches have been proposed. These methods are based on gene-specific data (either the raw data itself or $p$-values derived from gene-level statistical tests) overlayed on biological networks obtained

from databases such as Pathway Interaction Database [49], KEGG [50], Reactome [51], InnateDB [52], etc.

The relevance of network structure has long been appreciated. In [101], the authors presented systematic mathematical analysis of the topology of metabolic networks of 43 organisms representing all three domains of life, and found that despite significant variation in the pathway components, these networks share common mathematical properties which enhance error-tolerance. In [102], the authors compared the lethality of mutations in yeast with the positions of the affected protein in known pathways, and found that the biological necessity of the protein was well modeled by its connectivity in the network.

Based on such observations, Ideker et al. [103] proposed a method to identify subnetworks of pathways whose genes were enriched for highly significant genes. As the combinatorial problem of finding the maximum-scoring subnetwork is NP-hard (and hence computationally unfeasible for large networks), the authors introduced a simulated-annealing approximation. A related method, described in [104], searches for genes for which differential expression is present within the subnetwork of genes surrounding it. A more robust scoring approach improving upon [103] has recently been proposed [105], and is implemented in R/BioConductor as `BioNet` [106]. These techniques may be used to indicate subsets of interactions in a pathway that appear to be the most critical, and could be targeted in functional studies.

However, like non-network enrichment analyses (Section "Pathway Analysis"), these network-based enrichment analyses [102–104] rely upon the constituent genes displaying independent association with the phenotypes of interest and will fail to capture networks in which the individual gene expressions have similar distributions but altered coexpression characteristics. As an alternative, correlation- and co-expression-based approaches have been proposed in which the edges connecting two interacting genes are examined in the context of the surrounding network. In [107], the authors proposed an "activity" and "consistency" score for each interaction in a pathway. Beginning with a list of molecule input and outputs for every interaction in a biological pathway, Efroni et al. [107] defined the "activity" of the interaction as the joint probability of finding the interaction's genes in an overexpressed state and defined the "consistency" of the interaction as the probability of overexpression in the output conditioned on the activity of the inputs. Similarly, in ScorePAGE [108] the similarity between each pair of genes in a pathway is computed (e.g., correlation, covariance, etc.) and is averaged over the pathway weighted by the number of reactions needed to connect the two genes.

More recently, Signaling Pathway Impact Analysis (SPIA) [109, 110] was proposed. SPIA incorporates changes in gene expression with the types of interactions and the positions of genes in a pathway, defining a "perturbation factor" for each gene as the sum of its measured change in expression and a linear function of the perturbation factors of all the other genes in a pathway. Compared to GSEA [55], SPIA was found to have increased statistical power to detect altered pathways [110]. Similarly, the NetGSA method [111] also models each gene as a linear function of other genes in the network, but in addition accounts for a gene's baseline expression by representing it as a latent variable in the model. Both SPIA and NetGSA have been implemented as R BioConductor packages [13].

## *Network Inference*

In the methods outlined above, pathway network descriptions are obtained from curated databases and used as a framework in which to analyze transcriptomic data at the systems level. While this enables existing biological knowledge to be incorporated into the analysis, it also has the drawback of presuming that the network of regulatory relationships is accurately represented by the pathway database. A complementary approach involves the inference of regulatory networks from the data without making assumptions about the underlying graph. Network inference methods are thus able to identify previously unknown relationships between genes, as well as incorporate elements (such as microRNAs) that are not represented in pathway databases.

Inference of the underlying network structure given a set of cell states [112–120] is a formidable task. Although some success in the reconstruction of large-scale gene regulatory networks (GRNs) has recently been achieved in some cases [119–123], the systematic reconstruction of large-scale networks describing regulatory function and direct interactions of genes from expression or other data remains a major challenge in systems biology. With the increasing feasibility of genome-wide assays, an increasing amount of systems biology research is concerned with attempting to infer GRNs from large scale data sets based upon correlations between expression levels under various experimental conditions. Methods developed for this task are faced with a fundamental difficulty: while direct regulatory relationships between genes typically yield a high degree of correlation in their expression, the reverse is not necessarily true. For instance, two non-interacting genes may share the same upstream regulator, causing their expression to be correlated despite not sharing a direct link. On a global scale, GRNs are known to be sparse, i.e., direct regulatory relationships are a small fraction of all possible connections, but correlations can be non-vanishing between any pair of genes.

Increasingly sophisticated techniques have been devised to tackle this difficulty and attempt to infer the topological properties of GRNs from correlations in gene expression [112, 119]. Prominent examples are simple thresholding techniques [118], the use of partial correlation [116], and mutual information [113]. It should, however, be noted that an essential drawback of these methods is the reliance on arbitrary thresholds or related external parameters that are not defined by the system, and the quality of inferences based on these techniques often depends sensitively on these parameters. For instance, choosing correlation thresholds too high or too low yields false negatives or false positives, respectively.

Reconstructed networks may also be compared across phenotypes to identify novel interactions. In [124], the authors describe a method in which pairs of genes connected by a common edge in the pathway network were examined for correlation in tumor and normal gene expression data in multiple cancers. Gene–gene edges with correlations that exceeded a threshold were kept, thus forming a correlation network in tumors and a separate correlation network in normal cells. Differences in the resulting correlation networks were then assessed through a permutation test, indicating pathways with significant differences in gene correlation. (This method could be regarded as an network based extension of [62, 125, 126].)

Recently, these network inference techniques have played a role in reverse engineering the regulatory networks of healthy human B cells [127] and chronic lymphocytic leukemia cancer cells [128], providing a richer description of the systems biology of blood. Because network inference approaches do not rely on assumptions about the pathway architecture, they are exceptionally well-suited to be applied to integrated data sets (e.g., combining both mRNA and microRNA expression data) to identify complex regulatory relationships. In a recent example, network inference techniques have revealed how the networks of miRNAs and target genes are reprogrammed in leukemia [129], further enriching our understanding of the systems biology underlying healthy and diseased hematopoietic processes.

## Future Directions

Today, the feasibility of genome-wide assays, along with thousands of existing sequenced genomes [130] and hundreds of thousands of existing expression profiles [131] publicly available, provide exciting avenues for the investigation of developmental and disease processes in blood. To fully harness the power of this information, it is necessary not only to analyze the data at the gene-level, but also to examine it at the systems level. Driven by the abundance of experimental data, novel computational tools for systems-level investigations have been devised and implemented (including pathway enrichment analyses, methods for identifying functional gene-sets, and techniques for inferring regulatory networks), enabling a variety of complementary analytical techniques to be applied.

At the same time, a number of significant methodological challenges remain an area of active research, including improving the precision and accuracy of the knowledge contained in gene and pathways annotation databases, developing more efficient algorithms for combinatorially bound problems, and improving the robustness of network analysis and enrichment techniques. Just as the analytical methods will benefit experiment, so too will new experimental data inform methodological advances. We expect that these mutual advances will further improve the ability of computational and mathematical methods to model biological processes, predict clinical and experimental outcomes, and suggest therapeutic targets.

## References

1. van den Akker-van Marle ME, Gurwitz D, Detmar SB, Enzing CM, Hopkins MM, de Mesa EG, Ibarreta D. Cost-effectiveness of pharmacogenomics in clinical practice: a case study of thiopurine methyltransferase genotyping in acute lymphoblastic leukemia in Europe. Pharmacogenomics. 2006;7(5):783–92.
2. Karajannis M, Vincent L, Direnzo R, Shmelkov S, Zhang F, Feldman E, Bohlen P, Zhu Z, Sun H, Kussie P, Rafii S. Activation of fgfr1beta signaling pathway promotes survival, migration and resistance to chemotherapy in acute myeloid leukemia cells. Leukemia. 2006.

3. Savageau MA, Rosen R. Biochemical systems analysis: a study of function and design in molecular biology, vol. 725. Reading: Addison-Wesley; 1976.
4. Von Bertalanffy L: Modern theories of development: an introduction to theoretical biology. In: Woodger JH, transl. Oxford University Press; 1933 (originally published 1928).
5. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64.
6. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–93.
7. Parmigiani G. The analysis of gene expression data: methods and software. Springer; 2003.
8. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and snp calling from next-generation sequencing data. Nat Rev Genet. 2011;12(6):443–51.
9. Metzker ML. Sequencing technologies the next generation. Nat Rev Genet. 200911(1): 31–46.
10. Vazquez M, de la Torre V, Valencia A. Cancer genome analysis. PLoS Comput Biol. 2012 8(12):e1002824.
11. Smyth GK. Limma: linear models for microarray data. In: Bioinformatics and computational biology solutions using R and Bioconductor. Springer; 2005. pp. 397–420
12. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. http://www.R-project.org/. ISBN 3-900051-07-0.
13. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.
14. Gentleman, R., Carey, V., Huber, W., Irizarry, R., Dudoit, S.: Bioinformatics and computational biology solutions using R and Bioconductor, vol. 746718470. Springer; 2005.
15. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst. 2007;99(2):147–57.
16. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995; pp. 289–300.
17. Benjamini Y, Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001; pp. 1165–88.
18. Han, B., Kang, H.M., Eskin, E.: Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. PLoS Genet. 2009;5(4):e1000,456.
19. Csete ME, Doyle JC. Reverse engineering of biological complexity. Science 2002;295(5560), 1664–9.
20. Edelman GM, Gally JA. Degeneracy and complexity in biological systems. Proc Natl Acad Sci. 2001;98(24):13763–8.
21. D'haeseleer P. How does gene expression clustering work? Nat Biotechnol. 2005;23(12): 1499–501.
22. Datta S, Datta, S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics. 2003;19(4):459–66.
23. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci. 1998;95(25):14863–8.
24. Hartigan, J, Wong M. Algorithm AS 136: A $k$-means clustering algorithm. J R Stat Soc C Appl Stat. 1979;28:100–8.
25. Ng A, Jordan M, Weiss Y. On spectral clustering: analysis and an algorithm. Adv Neur Inf Process Syst. 2002;2, 849–56.
26. Leibon G, Pauls S, Rockmore D, Savell R. Topological structures in the equities market network. Proc Natl Acad Sci. 2008;105(52):20589–594.
27. Chung F. Spectral graph theory. American Mathematical Society; 1997.
28. von Luxburg U. A tutorial on spectral clustering. Stat Comput. 2007;17(4):395–416.

29. Qiu P, Plevritis SK. Simultaneous class discovery and classification of microarray data using spectral analysis. J Comput Biol. 2009;16:935–44.
30. Braun R, Leibon G, Pauls S, Rockmore D. Partition decoupling for multi-gene analysis of gene expression profiling data. BMC Bioinformatics. 2011;12(497).
31. Kim D, Lee K, Lee D. Detecting clusters of different geometrical shapes in microarray gene expression data. Bioinformatics 2005;21(9):1927–34.
32. Baker S. Simple and flexible classification of gene expression microarrays via swirls and ripples. BMC Bioinformat. 2010;11(1):452
33. Fraley C, Raftery A. MCLUST: Software for model-based cluster analysis. J. Classification 1999;16(2):297–306.
34. Still S, Bialek W. How many clusters? An information-theoretic perspective. Neural Comput. 2004;16(12):2483–506.
35. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc B. 2002;63(2):411–23.
36. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn. 2003;52 (1–2):91–118.
37. Monti S, Savage KJ, Kutok JL, Feuerhake F, Kurtin P, Mihm, M, Wu B, Pasqualucci L, Neuberg D, Aguiar RC, et al. Molecular profiling of diffuse large b-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. Blood. 2005;105(5):1851–61.
38. Jolliffe I. Principal component analysis. Wiley Online Library; 2005.
39. Wilson NK, Foster SD, Wang X, Knezevic K, Schütte J, Kaimakis P, Chilarska PM, Kinston S, Ouwehand WH, Dzierzak E, et al. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell. 2010;7(4):532–44.
40. Chambers SM, Boles NC, Lin KYK, Tierney MP, Bowman TV, Bradfute SB, Chen AJ, Merchant AA, Sirin O, Weksberg DC, et al. Hematopoietic fingerprints: an expression database of stem cells and their progeny. Cell Stem Cell. 2007;1(5):578–91.
41. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci. 2000;97(18):10101–6.
42. McIsaac RS, Petti AA, Bussemaker HJ, Botstein D. Perturbation-based analysis and modeling of combinatorial regulation in the yeast sulfur assimilation pathway. Mol Biol Cell 2012;23(15):2993–3007.
43. Narula J, Smith AM, Gottgens B, Igoshin OA. Modeling reveals bistability and low-pass filtering in the network module determining blood stem cell fate. PLoS Comput Biol. 2010;6(5):e1000771.
44. Bengio Y, Paiement J, Vincent P, Delalleau O, Le Roux N, Ouimet M. Out-of-sample extensions for LLE, IsoMap, MDS, eigenmaps, and spect ral clustering. Adv Neural Inf Process Syst. 2004;16:177–84.
45. Bengio Y, Delalleau O, Roux N, Paiement J, Vincent P, Ouimet M. Learning eigenfunctions links spectral embedding and kernel PCA. Neural Comput. 2004;16(10):2197–219.
46. Törönen P, Kolehmainen M, Wong G, Castrén E. Analysis of gene expression data using self-organizing maps. FEBS Lett. 1999;451(2):142–6.
47. Tamayo P, Slonim D, Mesirov J, Zhu Q, E Dmitrovsky SK, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci. 1999;96(6):2907–12.
48. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. Springer; 2009.
49. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009;37:D674–9.
50. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008;36(Database issue):D480–4.

51. Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L. Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 2007;8(3):R39.
52. Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, Barsky A, Gardy JL, Roche FM, Chan TH, Shah N, et al. Innatedb: facilitating systems-level analyses of the mammalian innate immune response. Mol Syst Biol. 2008;4(1).
53. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. BioMart–biological queries made easy. BMC Genomics. 2009;10:22.
54. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.
55. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43):15545–50.
56. Jiang Z, Gentleman R. Extensions to gene set enrichment. Bioinformatics. 2007;23(3): 306–13.
57. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007;23(8):980–7.
58. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci U S A. 2005;102(38):13544–9.
59. Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R. Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. Blood. 2004;103(7):2771–8.
60. Grigoryev YA, Kurian SM, Avnur Z, Borie D, Deng J, Campbell D, Sung J, Nikolcheva T, Quinn A, Schulman H, et al. Deconvoluting post-transplant immunity: cell subset-specific mapping reveals pathways for activation and expansion of memory t, monocytes and b cells. PloS One. 2010;5(10):e13,358.
61. Ma S, Kosorok MR. Identification of differential gene pathways with principal component analysis. Bioinformatics. 2009;25(7):882–9.
62. Braun R, Cope L, Parmigiani G. Identifying differential correlation in gene/pathway combinations. BMC Bioinformatics. 2008;9:488.
63. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to dna microarrays. Stat Sci. 2003;104–17.
64. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286(5439):531–7.
65. Hastie T, Tibshirani R, Narasimhan B, Chu G. pamr: Pam: prediction analysis for microarrays. 2011. http://CRAN.R-project.org/package=pamr. R package version 1.54.
66. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 2005;6(2):95–108.
67. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9(5):356–69.
68. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci. 2009;106(23):9362–7.
69. Schork N, Murray S, Frazer K, Topol E. Common vs. rare allele hypotheses for complex diseases. Current Opin Genet Dev. 2009;19(3):212–9.
70. Moore J, Asselbergs F, Williams S. Bioinformatics challenges for genome-wide association studies. Bioinformatics. 2010;26(4):445.
71. Greene C, Penrod N, Williams S, Moore J. Failure to replicate a genetic association may provide important clues about genetic architecture. PLoS One. 2009;4(6):e5639.

72. Moore J. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered. 2003;56(1–3):73–82.
73. Tyler A, Asselbergs F, Williams S, Moore J. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. BioEssays. 2009;31(2):220–7.
74. Holmans P. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. Adv Genet. 2010;72:141.
75. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet. 2010;11(12):843–54.
76. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet. 2007;81(6):1278.
77. Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics. 2008;24(23):2784–5.
78. Motsinger A, Ritchie M. Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene–gene interactions in human genetics and pharmacogenomics studies. Hum Genomics. 2006;2(5):318–28.
79. Moore J, Gilbert J, Tsai C, Chiang F, Holden T, Barney N, White B. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol. 2006;241(2):252–61.
80. Cordell H. Detecting gene–gene interactions that underlie human diseases. Nat Rev Genet. 2009;10(6):392–404.
81. Greene C, Sinnott-Armstrong N, Himmelstein D, Park P, Moore J, Harris B. Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic als. Bioinformatics. 2010;26(5):694.
82. Kira K, Rendell L. A practical approach to feature selection. Proceedings of the Ninth International Workshop on Machine learning; 1992. pp. 249–56.
83. Robnik-Šikonja M, Kononenko I. An adaptation of relief for attribute estimation in regression. Proceedings of the International Conference on Machine Learning ICML-97; 1997. pp. 296–304.
84. Moore J. Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics. Knowledge Discovery and Data Mining: Challenges and Realities with Real World Data; 2007. pp. 17–30.
85. Greene C, Penrod N, Kiralis J, Moore J. Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. BioData Mining. 2009;2:5.
86. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008;4(8):e1000167.
87. Braun R, Rowe W, Schaefer C, Zhang J, Buetow K. Needles in the haystack: Identifying individuals present in pooled genomic data. PLoS Genet. 2009;5(10):e1000668.
88. Visscher PM, Hill WG. The limits of individual identification from sample allele frequencies: theory and statistical analysis. PLoS Genet. 2009;5(10):e1000628.
89. Braun R, Buetow K. Pathways of Distinction Analysis: a new technique for multi-SNP ana lysis of GWAS data. PLoS Genet. 2011;7(6):e1002101.
90. Breiman L. Random forests. Machine Learn. 2001;45(1):5–32.
91. Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, Floyd E, Zhao H. Pathway analysis using random forests classification and regression. Bioinformatics. 2006;22(16):2028–36.
92. Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006;7(1):3.
93. Dettling M. Bagboosting for tumor classification with gene expression data. Bioinformatics. 2004;20(18):3583–93.
94. Lee JW, Lee JB, Park M, Song SH. An extensive comparison of recent classification tools applied to microarray data. Comput Stat Data Anal. 2005;48(4):869–85.
95. Hassane DC, Guzman ML, Corbett C, Li X, Abboud R, Young F, Liesveld JL, Carroll M, Jordan CT. Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data. Blood. 2008;111(12):5654–62.

96. Van Ness B, Ramos C, Haznadar M, Hoering A, Haessler J, Crowley J, Jacobus S, Oken M, Rajkumar V, Greipp P, et al. Genomic variation in myeloma: design, content, and initial application of the bank on a cure snp panel to detect associations with progression-free survival. BMC Med. 2008;6(1):26.
97. Ackermann M, Sikora-Wohlfeld W, Beyer A. Elucidating the regulatory mechanisms of transcription factor activity in hematopoietic stem cell differentiation. In: Saxon Biotechnology Symposium; 2011. p. 79.
98. De Souto M, Costa I, De Araujo D, Ludermir T, Schliep A. Clustering cancer gene expression data: a comparative study. BMC Bioinformatics. 2008;9(1):497.
99. Kolaczyk ED. Statistical analysis of network data. Springer; 2009.
100. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101–13.
101. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. Nature. 2000;407(6804):651–4.
102. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001;411(6833):41–2.
103. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science. 2001;292(5518):929–34.
104. Nacu S, Critchley-Thorne R, Lee P, Holmes S. Gene expression network analysis and applications to immunology. Bioinformatics. 2007;23(7):850–8.
105. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. Bioinformatics. 2008;24(13):i223–31.
106. Beisser D, Klau GW, Dandekar T, Müller T, Dittrich MT. Bionet: an r-package for the functional analysis of biological networks. Bioinformatics. 2010;26(8):1129–30.
107. Efroni S, Schaefer CF, Buetow KH. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. PLoS One. 2007;2(5):e425.
108. Jörg R, Jochen M, Thomas L, et al. Calculating the statistical significance of changes in pathway activity from gene expression data. Stat Appl Genet Mol Biol. 2004;3(1):1–31.
109. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R. A systems biology approach for pathway level analysis. Genome Res. 2007;17(10):1537–45.
110. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim Js, Kim CJ, Kusanovic JP, Romero R. A novel signaling pathway impact analysis. Bioinformatics. 2009;25(1):75–82.
111. Shojaie A, Michailidis G. Penalized principal component regression on graphs for analysis of subnetworks. In: Advances in neural information processing systems; 2010. pp. 2155–63.
112. Bansal M, Belcastro V, Ambesi-Impiombato A, Di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol. 2007;3(1).
113. Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006;7(Suppl 1):S7.
114. Gardner T, Faith J. Reverse-engineering transcription control networks. Phys Life Rev. 2005;2(1):65–88.
115. Meyer P, Lafitte F, Bontempi G. minet: An R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics. 2008;9(1):461.
116. de la Fuente A, Brazhnik P, Mendes P. Linking the genes: inferring quantitative gene networks from microarray data. TRENDS Genet. 2002;18(8);395–8.
117. Gardner T, di Bernardo, D, Lorenz D, Collins J: Inferring genetic networks and identifying compound mode of action via expression profiling. Sci Signal. 2003;301(5629):102.
118. Rice J, Tu Y, Stolovitzky G. Reconstructing biological networks using conditional correlation analysis. Bioinformatics. 21(6):765–73.
119. Marbach D, Prill R, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G: Revealing strengths and weaknesses of methods for gene network inference. Proc Natl Acad Sci. 2010;107(14):6286–91.

120. Altay G, Emmert-Streib F: Revealing differences in gene network inference algorithms on the network level by ensemble methods. Bioinformatics. 2010;26(14):1738–44.
121. Dodd IB, Micheelsen MA, Sneppen K, Thon G. Theoretical analysis of epigenetic cell memory by nucleosome modification. Cell. 2007;129(4):813–22.
122. Sedighi M, Sengupta AM. Epigenetic chromatin silencing: bistability and front propagation. Phys Biol. 2007;4(4):246–55.
123. Graf T, Enver T. Forcing cells to change lineages. Nature. 2009;462:(7273):587–94.
124. Choi JK, Yu U, Yoo OJ, Kim S. Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics. 2005;21(24):4348–55.
125. Ho YY, Cope L, Dettling M, Parmigiani G. Statistical methods for identifying differentially expressed gene combinations. In: Gene function analysis. Springer; 2007. pp. 171–91.
126. Dettling M, Gabrielson E, Parmigiani G. Searching for differentially expressed gene combinations. Genome Biol. 2005;6:R88.
127. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human b cells. Nat Genet. 2005;37(4):382–90.
128. Vallat L, Kemper CA, Jung N, Maumy-Bertrand M, Bertrand F, Meyer N, Pocheville A, Fisher JW, Gribben JG, Bahram S. Reverse-engineering the genetic circuitry of a cancer cell with predicted intervention in chronic lymphocytic leukemia. Proc Natl Acad Sci. 2013;110(2):459–64.
129. Volinia S, Galasso M, Costinean S, Tagliavini L, Gamberoni G, Drusco A, Marchesini J, Mascellani N, Sana ME, Jarour RA, et al. Reprogramming of miRNA networks in cancer and leukemia. Genome Res. 2010;20(5):589–99.
130. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2011;39(Suppl 1):D38–51.
131. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009;37(Suppl 1):D885–90.