Mining Massive Genomic Data for Therapeutic Biomarker Discovery in Cancer: Resources, Tools, and Algorithms

Pan Tong and Hua Li

Abstract Cancer research is experiencing an evolution empowered by high-throughput technologies that makes it possible to collect molecular information for the entire genome at the DNA, RNA, protein, and epigenetic levels. Due to the complex nature of cancer, several organizations have launched comprehensive molecular profiling for thousands of cancer patients using multiple high-throughput technologies to investigate cancer genomics, transcriptomics, proteomics, and epigenomics. To speed up the bench-to-bedside translation, additional efforts have been made to profile hundreds of preclinical cell line models coupled with systematic screening of anticancer agents. This leads to an explosion of massive genomic data that shifts the bottleneck from data generation to data analytics. In this chapter, we will first introduce different types of genomic data as well as resources from publicly accessible data repositories that can be utilized to search for therapeutic targets for cancer treatment. We then introduce software tools frequently used for genomic data mining. Finally, we summarize working algorithms for the discovery of therapeutic biomarkers.

Keywords Genomics • Transcriptomics • Proteomics • Epigenomics • Biomarker discovery • Cancer

1 Introduction

Cancer is a disease of genetics involving dynamic changes of the genome [1]. Multiple genetic alterations have been identified in cancer including somatic mutations, DNA copy number change, epigenetic modifications, and dysregulated gene expression. Systematic discovery of cancer-driven alterations not only helps

P. Tong
Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

H. Li (✉)
School of Biomedical Engineering, Bio-ID Center, Shanghai Jiao Tong University, Shanghai, China
e-mail: kaixinsjtu@hotmail.com

K.-C. Wong (ed.), *Big Data Analytics in Genomics*,
DOI 10.1007/978-3-319-41279-5_10

us better understand tumorigenesis but also plays crucial roles in developing biomarkers for cancer detection, diagnosis, and prognosis. Over the last decade, there has been a dramatic advance in technologies allowing holistic interrogations of various aspects of cellular process including mRNAs (transcriptome), proteins (proteome), sequence and structural variations (genomics), metabolites (metabolomics), and interactions (interactome). While data from individual assay type is informative for certain aspects of biology, integrative analysis using multi-assay data sets is more powerful and provides deeper insights into understanding complex biological systems and diseases. As a result, there is an increasing trend for both individual laboratories and large consortia to generate multi-assay genomic profiling of cancer patients. For example, pioneering studies from the Sanger Institute and the Johns Hopkins Hospital identified frequent mutations in melanoma and colon cancer [2, 3]. Later studies in Boston and New York uncovered activating mutations in lung cancer which predicted response to kinase inhibitors [4–6]. Soon thereafter, the Human Cancer Genome Project was proposed by US National Cancer Institute which was later called The Cancer Genome Atlas (TCGA) [7]. In parallel, the International Cancer Genome Consortium (ICGC) was launched to foster international collaborations for large-scale cancer genomics studies [8]. To speed up the transition from bench to bedside, cell line models derived from cancer patients have been under extensive investigation. Several studies have generated comprehensive genomic characterizations of hundreds of cell line models coupled with drug screening enabling us to generate predictors of drug sensitivity based on genomic information [9–12].

While genomic data is now generated faster and cheaper than ever before, our ability to manage, analyze, and interpret it has not paced with the data deluge. Consequently, for the first time in history, the bottleneck in cancer research is shifting from data collection to data mining [13]. The objective of this chapter is to bridge the gap between advances in high-throughput genomics and our ability to manage, integrate, and analyze genomic data with special focus on therapeutic biomarker discovery. We begin with the definition of biomarker and an overview of different types of genomic data. We then summarize resources from publicly data repositories that can be utilized to search for therapeutic targets. We further introduce software tools frequently used for genomic data mining. Finally, we discuss working algorithms for the discovery of biomarkers.

2 Biomarkers in Cancer

According to National Cancer Institute, a biomarker is defined as a molecule found in blood or other body fluids that is objectively measured and evaluated as an indicator of disease status, pathogenic processes, or pharmacologic responses to therapeutic agents [14]. Biomarkers have been utilized for various applications including (1) measuring the natural history of disease, (2) correlating with clinical outcomes, (3) determining the biological effect of a therapeutic intervention, and

(4) serving as surrogate endpoints in clinical trials [15, 16]. Based on their utility, several types of biomarkers exist. Diagnostic biomarkers are used for early disease detection. Predictive biomarkers can infer the efficacy or toxicity of a drug. Prognostic biomarkers are used to assess if a patient receiving treatment will perform well or whether more aggressive treatment is needed to prevent recurrence. Staging biomarkers are used to determine the stage of progression of a disease.

Several methods can be used to identify candidate biomarkers. The classic approach is to identify biomarkers based on tumor biology where pivotal molecules in regulatory pathways are selected as candidates. However, such approach is time-consuming giving the large number of molecules and metabolites to be searched for. Recent development of high-throughput technologies has brought biomarker discovery into the "omics" era enabling simultaneous measurement of thousands of molecules. Genomics studies involving genotyping and next-generation sequencing have identified a considerable amount of biomarkers (such as single-nucleotide polymorphisms and structural variations) associated with drug efficacy and disease progression [17, 18]. Similarly, transcriptome and proteome profiling have also revealed biomarkers (such as dysregulated expression of RNA and proteins) that are highly correlated with clinical outcomes [19, 20].

3 High-Throughput Genomics

Over the past decade, there has been a dramatic advance in technologies enabling genome-scale data collection regarding various aspects of cellular process including sequence and structural variation, transcriptome, epigenome, and proteome. These technologies generate massive amounts of genomic data faster than ever before. Below we summarize major types of genomics data in cancer research and related technologies used to collect such data.

3.1 Transcriptomics

Transcriptomics is the study of the complete set of RNA transcripts produced by the genome. Data collected for transcriptome starts with DNA microarrays using either spotted oligonucleotides or in situ synthesized probes to quantitatively measure mRNA levels of a large number of genes. The emergence of low-cost short read sequencing, also known as next-generation sequencing (NGS) technology, escalates transcriptomics studies to a new level by overcoming many drawbacks inherent in microarray such as requirement of carefully designed probes, cross-hybridization, high background noise, and low resolution [21]. In addition to provide fast and accurate measurement of transcripts, NGS RNA sequencing also facilitates deeper understanding of the transcriptome including alternative splicing, gene fusion, and isoform expression. It is worth noting that transcriptomics studies are not limited

to the investigation of messenger RNA. For example, whole genome profiling of microRNAs and other noncoding RNAs is usually employed to decipher post-transcriptional regulation of gene expression [22].

3.2 Proteomics

Proteomics is the large-scale study of proteins including protein abundance, modifications, localizations and interactions. The growth of proteomics studies owes to the advances in protein technologies such as capillary electrophoresis, high performance liquid chromatography (HPLC), matrix-assisted laser desorption/ionization (MALDI), and mass spectrometry [23]. The reverse-phase protein array (RPPA) first introduced in the early 2000s is widely used in biomarker discovery, therapeutic target evaluation, and cancer research. It now becomes a promising tool for clinical trials [24].

3.3 Epigenomics

Epigenomics refers to the study of epigenetic modifications in the DNA sequence as well chromatin including DNA methylation, covalent modifications of cytosine, and post-translational modifications of histones such as methylation, acetylation, and phosphorylation [25]. Functionally, epigenetic modifications are involved in regulation of gene expression, gene dosage, chromosome inactivation, and genome imprinting. It has been found that changes in epigenomics have been implicated in multiple diseases including cancer [26]. Epigenomics can be studied using DNA methylation array or next-generation sequencing with chemically treated DNA [27].

3.4 Sequence Variation

Genomic sequence variation includes single-nucleotide polymorphisms (SNP), mutations, copy number variations, and structural variations. The ultimate goal of human genetics is to identify all genomic sequence variation and deciphers how they contribute to phenotype and diseases. Currently, SNP arrays are cost-effective instruments to identify SNPs and copy number variations. In contrast, NGS technologies can be applied to interrogate all the genomic variations and provide higher resolution data for downstream functional studies [28].

4 Resources for Genomic Data

There is a rich source of public genomic data which provides unprecedented possibilities for hypothesis generation and data mining.

4.1 Gene Expression Omnibus (GEO)

GEO (http://www.ncbi.nlm.nih.gov/geo) is the largest public repository for high-throughput gene expression data [29]. GEO archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data generated by the scientific community. There are three main entities in GEO database: Platform, Sample, and Series [30]. A Platform record includes a summary description of the array or sequencer and an additional table providing probe annotation or sequence information. Each Platform record is assigned a unique GEO accession number with prefix GPL. A Sample record provides all information related to a sample including phenotype information, experimental protocol, and abundance measurements for each feature recorded in the Platform. Sample accession numbers have a GSM prefix. A Series record defines a set of Samples related to a particular study and provides a description of overall study design. Series records have a prefix GSE. As of 2013, the GEO database hosts >32,000 records of Series submitted by around 13,000 laboratories, corresponding to 800,000 samples derived from over 1600 organisms [31]. Genomic data is worthless without contextual biological details and analysis methodologies for preprocessing. To ensure important information is preserved, scientific reporting standards have been proposed such as MIAME (Minimum Information About a Microarray Experiment) for data annotation and MINiML (MIAME Notation in Markup Language) for XML based data exchange [31]. The GEO database is in compliance with both MIAME and MINiML standards which greatly facilitates data retrieval. In addition to provide a searchable database for data retrieval, GEO now includes basic data mining and visualization functionalities. Users can compare two sets of samples with specified statistical parameters, construct clustered heatmaps, retrieve profiles with similar patterns of expression, and identify profiles belonging to the same homologs [32]. A major update recently was the release of GEO2R web application which allows users to perform sophisticated analysis using R [31]. Once a user specifies a Series number to be analyzed, GEO will retrieve the data using GEOquery [33] in the backend. The retrieved data is then subjected to analysis specified by user or from default settings. Results computed from the server are transferred to user using JSON and rendered as HTML page. Since R script is provided, users can always reproduce the analysis and fine-tune it offline.

4.2 ArrayExpress

ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) is the European counterpart of GEO. ArrayExpress complies with the MIAME and MAGE-ML (Microarray Gene Expression Markup Language) standards to ensure data consistency. Currently, ArrayExpress contains more than 1.8 million samples with high-throughput assays across over 62,000 studies with a total file size of 40 TB. Programmatic access of the ArrayExpress data is available through the ArrayExpress Bioconductor package [34].

4.3 The Cancer Genome Atlas (TCGA)

The first public repository dedicated to cancer is TCGA (https://tcga-data.nci.nih.gov/tcga/). The overall goal of TCGA is to generate comprehensive, multi-dimensional profiling of genomic alterations in major cancer types. TCGA is organized by different centers responsible for sample collection, processing, and analysis. First, Tissue Source Sites (TSSs) collect biospecimens from eligible cancer patients and deliver them to Biospecimen Core Resources (BCRs). BCRs then catalogue, process, and verify the received samples before submitting to Data Coordinating Centers (DCC). DCCs provide molecular analytes for the Genome Characterization Centers (GCCs) and Genome Sequencing Centers (GSCs) for genomic characterization. The generated genomic data is passed to Genome Data Analysis Centers (GDACs) for information processing, analysis, and tool development. All data generated is made publicly available through TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp) and CGHub (https://cghub.ucsc.edu/). TCGA employs several high-throughput technologies including microarrays, next-generation sequencing, and reverse-phase protein array (RPPA) to interrogate global alterations at DNA, RNA, and protein levels. In particular, RNA sequencing provides transcriptomic monitoring of gene expression, isoforms, gene fusions, and noncoding RNAs. DNA sequencing determines genetic alterations such as insertions, deletions, polymorphisms, and copy number variations. SNP-based platforms assess single-nucleotide polymorphisms (SNPs), copy number variations, and loss of heterozygosity (LOH). Array-based methylation provides epigenetic information at CpG sites. Bisulfite sequencing characterizes DNA methylation at single nucleotide resolution. RPPA provides quantitative measurements of protein expression with high sensitivity. Since its inaugural in 2006, TCGA has comprehensively profiled more than 10,000 samples across 33 cancer types.

4.4 International Cancer Genome Consortium (ICGC)

While TCGA provides comprehensive genomic characterization for cancer patients in the USA, the ICGC project (https://icgc.org/) aims to generate an extensive catalogue of genomic abnormalities for cancer patients throughout the world contributed by different participating countries. Currently, the ICGC data portal records 78 projects covering 50 different cancer types. The ICGC data portal periodically updates with newly generated data and provides tools for data downloading, visualizing, and querying. Due to the large size of data which may take months to download, ICGC partners with Amazon Web Services to facilitate data access through the cloud. ICGC also releases analytic workflows so that users can analyze their data using the same workflows after initiating an Amazon machine.

4.5 The NCI-60 Cell Line Panel

Immortalized cell lines derived from human cancer have made significant contributions to cancer biology and formed the basis of current understanding of drug sensitivity and resistance. Therefore, systematic genomic characterization of cell line models coupled with pharmacological interrogation would greatly facilitate biomarker identification and drug development. One of the early endeavors is the NCI-60 project which has released a panel of 60 cell lines with high-throughput gene genomic profiling including DNA copy number, gene expression, protein expression, and mutation and additional anticancer drug screening [11]. The NCI-60 panel quickly becomes a rich source of information to investigate mechanisms of drug resistance. A major discovery using the NCI-60 data set has been the linkage between P-glycoprotein expression and multi-drug resistance [35].

4.6 Cancer Cell Line Encyclopedia (CCLE)

Following the success of NCI60, the CCLE project (http://www.broadinstitute.org/ccle/home) [9] has extended genomic characterization to around 1000 cell lines using gene expression, chromosomal copy number, and massively parallel sequencing technology. CCLE has also screened 24 anticancer drugs from 479 cell lines using an automated compound-screening platform [9]. An integrative analysis of the CCLE data identified genetic, lineage and gene expression based predictors of pharmacological vulnerabilities which had important clinical implications for personalizing cancer therapy [9].

4.7 Genomics of Drug Sensitivity in Cancer (GDSC)

Similar to CCLE, the GDSC (http://www.cancerrxgene.org/) database has profiled 138 anticancer drugs encompassing both targeted agents and cytotoxic therapeutics across 700 cancer cell lines [12]. Initial analysis using GDSC data found that mutated cancer genes were markers of sensitivity or resistance to a broad range of anticancer drugs. Further, the mutated cancer genes mostly associated with sensitivity were found to be oncogenes that were direct targets of the drug [12]. On the other hand, inactivating mutations in tumor suppressors were associated with drug resistance [12]. For example, mutations in BRAF, an oncogene responsible for protein kinase signaling, were associated with sensitivity of BRAF inhibitors MEK1 and MEK2. In contrast, mutations in TP53, an important tumor suppressor responsible for apoptosis, conferred resistance to nutlin-3a, which was an inhibitor of MDM2 that negatively regulated p53 protein [36].

4.8 Cancer Therapeutics Response Portal (CTRP)

In addition to identify biomarkers of drug sensitivity, genomic characterization coupled with drug screening can also shed light on mechanisms of action (MoA). Recently, the CTRP (http://www.broadinstitute.org/ctrp/) database published high quality screening data of 481 compounds across 860 cancer cell lines spanning 23 lineages [10]. By comparing the sensitivity pattern of compounds targeting the same gene, targeting genes in the same pathway and targeting genes that metabolically process the compounds, the authors observed that sensitivity may depend on metabolic activation, import of the compound, the presence of target-drug complex, and the presence of target expression. On the other hand, drug resistance was linked to drug inactivation or an efflux mechanism that depleted drug from the cell [10].

4.9 Project Achilles

In an effort to identify genes essential for cell proliferation and viability in cancer cell lines, Project Achilles (https://www.broadinstitute.org/achilles) employs genome-wide genetic perturbation experiments using pooled shRNA technology. The screening pipeline uses around 54,000 shRNA plasmids targeting 11,000 genes with a minimum representation of 200 cells per shRNA [37]. The pooled shRNA screens are able to silence or knock-out genes and thus identify genes essential for growth and survival. After incubation for a certain period of time, the cell lines are harvested to determine relative levels of shRNA plasmids using Illumina sequencing technology. When linked with genetic characteristics of the cell lines, Project Achilles provides valuable information for prioritizing targets for therapeutic drug development.

While individual data resource introduced here can be helpful in addressing different questions, it is usually more valuable to integrate across different resources since they largely provide complementary information. For example, candidate biomarkers overexpressed in cancers can be identified using the TCGA data. The therapeutic relevance of such biomarkers in terms of in vitro drug sensitivity can then be evaluated using the NCI-60 panel, the CCLE, and the GDSC database. Finally, essentiality of these biomarkers from knock-out experiments can be extracted from the Project Achilles data. Such an integrated analysis not only provides a full picture of the utility associated with identified biomarkers, but greatly narrows down the number of candidates and thus can greatly reduce costs in validating the biomarkers.

5 Tools for Mining Genomic Data

Choosing the right set of tools is vital for genomic data mining. One of the most popular tools is the R programming language, an open source environment for statistical computing. R has strong support for statistical analysis including linear and nonlinear modeling, hypothesis testing, time series analysis, spatial analysis, clustering, and classification. R also provides various facilities for data manipulation, calculation, and visualization [38]. Further, R is highly extensible with lots of packages contributed by users in the R community. Among the various packages dedicated to high-throughput genomics, Bioconductor is one of the most comprehensive and versatile tools [39]. It greatly facilitates rapid creation of pipelines by combining multiple procedures. Bioconductor includes tools for all stages of analysis ranging from data generation to final presentation. Bioconductor also has high quality documentation through three levels: vignettes that provide example usages of a particular package; manual pages that precisely describe inputs, outputs and examples of a function; and workflows that showcase complete analysis spanning multiple tools and packages. Recently, Bioconductor has enhanced its functionality by enabling analysis of next-generation sequencing data. Core infrastructure includes *Biostrings* for DNA and amino acid sequence manipulation, *ShortRead* for FASTQ files, *IRanges* and *GenomicRanges* for genome coordinate calculation, *GenomicAlignments* and *Rsamtools* for aligned sequencing data, *BSgenome* for curated whole-genome sequence, and *rtracklayer* for integration of genome browsers with experimental data. Currently, Bioconductor has 1104 packages, 895 annotation databases, and 257 packaged experimental data and the functionality is still expanding.

Genomic data will be useless if no metadata is given regarding the entities measured such as gene symbols, probe ID, genomic coordinates, and genome versions. Public service providers and instrument vendors have websites from which users can download relevant information for offline data wrangling. However, this process is time-consuming, error-prone, and irreproducible. The *biomaRt* package hosted on the Bioconductor repository provides a unified interface for accessing a

large collection of databases including NCBI (National Center for Biotechnology Information), Ensembl, UCSC (University of California, Santa Cruz), COSMIC (Catalogue of Somatic Mutations in Cancer), Uniprot (Universal Protein Resource), HGNC (HUGO Gene Nomenclature Committee providing official gene names), and Reactome (curated biological pathways) [40]. *BiomaRt* allows seamless integration of identifier mapping and annotation into data analysis, creating a powerful platform for biological data mining [41].

Although R combined with Bioconductor proves to be a powerful computing engine for genomic analysis, users are required to have reasonable programming skills to fully unleash its power. Alternately, there are web-based tools suited for both experimentalists and computational colleagues where analysis can be performed with mouse click. Galaxy is one of such tools with a web-based graphical user interface for accessible, reproducible, and transparent genomic data mining [42]. By encapsulating high-end computation tools while hiding the technical details of computation and storage, the Galaxy software becomes highly accessible to users without programming skills. By automatically tracking metadata regarding input data sets, analysis parameters, analytic components and output data, and by supporting user specified annotations and tags, Galaxy makes it easy to assemble and reproduce any given analysis [43]. Galaxy also makes analysis transparent by allowing users to share their analysis using Galaxy's sharing model. This includes a web-based framework for displaying results, customizable web pages that users

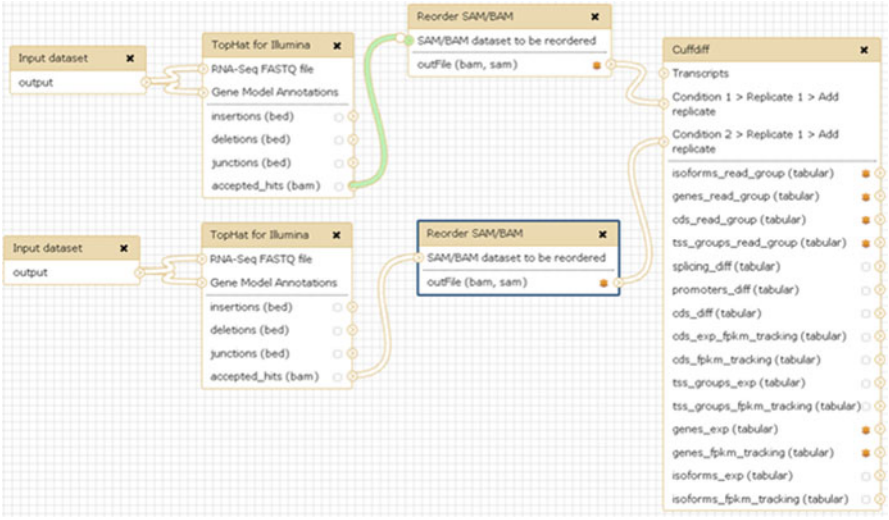

Fig. 1 An example Galaxy workflow for RNAseq differential expression analysis. Each box represents a tool with input and output files. Users can connect the output files of one component to the input of another component to form a complete analysis. A compatible link between components will become green to aid users. Here input FASTQ files are fed into Tophat for alignment. The resulting bam file is then sorted and indexed for Cuffdiff differential gene expression analysis

can freely modify, and a public repository hosting published items such as datasets, histories, and workflows [42].

The Galaxy workflow greatly enhances usability by providing a drag-and-drop interface for building analytic pipelines. Figure 1 demonstrates a workflow for RNAseq differential expression analysis. Users insert different analytic components into the workflow canvas and connect them to form a complete analysis. The workflow editor verifies each link between the tools for compatibility. Compatible links will turn green transiently to visually aid workflow construction. To further simplify the creation of workflows, Galaxy allows users to create workflows from analysis history. This feature greatly simplifies workflow usage since users do not need to plan analysis upfront. In addition, Galaxy is highly extensible. Any piece of software written in any language can be integrated into the Galaxy workflow. To add a new tool to Galaxy, users only need to specify a configuration file dictating how to run the tool. Additionally, users need to describe input and output parameters so that the Galaxy framework knows how to work with the tool abstractly and automatically generates a web interface for it.

An alternative tool to Galaxy is the GenePattern software (www.genepattern.org/) which also does not require programming skills. GenePattern is claimed as a pipeline builder providing form-based methods for data preprocessing, analysis, and visualization [44]. GenePattern hosts different modules through a centralized repository so that users can download or upgrade when needed. In addition to the graphical user interface, GenePattern also allows command line access which makes automatic batch processing possible. Currently, users can access GenePattern through R, Matlab, and Java by invoking a local GenePattern instance. The combination of a graphical user interface with a programmatic console becomes a unique feature of GenePattern. Since its first release in 2004, GenePattern has over 23,000 registered users from over 2900 commercial and non-profit organizations worldwide.

There are also tools built on well-curated cancer genomic data. Here we illustrate two examples: cBioPortal (http://www.cbioportal.org/) and Oncomine (http://www.oncomine.org). cBioPortal provides a web interface for exploring, visualizing, analyzing, and downloading multi-platform cancer genomic data [45]. By hosting a large set of well-curated cancer genomic data including somatic mutation, mRNA and microRNA expression, protein expression, DNA copy number, and DNA methylation, cBioPortal greatly facilitates integrative genomic analysis by allowing users to query multiple data types and their associations at individual gene level. Further, cBioPortal also supports mutual exclusivity analysis for genomic alterations, survival analysis, co-expression analysis, enrichment analysis, and network analysis. While cBioPortal focuses on multi-platform cancer genomic data, Oncomine specializes in microarray gene expression data. Currently, Oncomine hosts 715 datasets consisting of 86,733 samples. Oncomine also provides a web interface for users to perform differential gene expression, co-expression, interaction network, cancer outlier profile analysis (COPA), and molecular concept analysis [46].

6 Algorithms for Cancer Biomarker Discovery

To identify biomarkers of various utilities, both supervised and unsupervised methods can be used. For supervised biomarker discovery, an outcome variable associated with each sample is required so that candidate markers predictive to the outcome variable can be identified. In comparison, unsupervised methods rely on the genomic assays only and search candidate biomarkers by modeling the signals from genomic measurements. Below we introduce several popular algorithms for both supervised and unsupervised biomarker discovery.

6.1 Supervised Methods

A common task of genomic biomarker discovery is to compare the gene expression levels (e.g., transcriptomic expression, proteomic expression, or microRNA expression) of samples under different treatment conditions or at different time points. This task is usually called a differential gene expression (DEG) analysis. Numerous methods have been published for DEG analysis using high-throughput genomics data. A straightforward approach is to use a two-sample *t*-test (in the case of binary outcome) or a linear regression framework (in the case of categorical or continuous outcome). However, genomic data may contain outlier measurements that violate the underlying statistical assumptions. Therefore, various variants of these methods have been developed by considering statistical robustness. One of the most popular methods used in gene expression analysis is the significance analysis of microarray (SAM) software developed by Tusher et al. [47]. For a two-sample comparison, SAM computes a "relative difference" metric $d(i)$ for each gene i:

$$d(i) = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s(i) + s_0}$$

Here $\bar{x}_1(i)$ and $\bar{x}_2(i)$ are the average expression in the two groups, $s(i)$ is the gene specific standard deviation of the repeated measurements, and s_0 is a stabilizing constant chosen to penalize uninteresting genes with poor signal to noise ratio. Since a theoretical null distribution of $d(i)$ is difficult to obtain, SAM instead resorts to a permutation based approach to assess statistical significance.

The original SAM software only dealt with binary outcome. Later versions of SAM allowed the analysis of data with multiclass outcome, continuous outcome, and censored survival time. To do this, the authors extended the definition of $d(i)$ as following:

$$d(i) = \frac{r_i}{s(i) + s_0}$$

Here r_i is a score defined differently for different types of outcome. More details about this extension can be found on the SAM manual (http://statweb.stanford.edu/~tibs/SAM/).

An important concept for biomarker discovery is that statistical significance does not ensure biological significance. Clinically useful biomarkers need to have strong dynamic range and a manageable gene size so that they can be easily assayed in a single panel. The SAM method computes a score for each gene and additional filtering step is needed to narrow down the gene list. The Top Scoring Pair (TSP) method uses a different strategy by comparing the relative expression of every possible gene pairs [48]. It effectively reduces the number of biomarkers to two and limits the selected genes to have a strong contrast easy to quantify [48]. Let the expression of a particular gene in sample i be x_i and let the class label associated with each sample be c which can be any value in $\{1, 2, \ldots, C\}$. TSP computes the frequency of observing $x_i < x_j$ for each class c as $p_{ij}(\mathrm{c}) = \mathrm{P}\left(x_i < x_j \middle| \mathrm{c}\right)$. In the case of $\mathrm{C} = 2$, the TSP score Δ_{ij} is defined as following although this metric can be extended to higher number of classes:

$$\Delta_{ij} = \left| p_{ij}(1) - p_{ij}(2) \right|$$

The TSP method selects genes based on their relative expression, which is different from other approaches used in DEG analysis. Further, a TSP pair provides a simple rule to classify samples into different classes. For example, if gene i has higher expression than gene j in a TSP pair, the sample will be classified as class 1 or class 2 otherwise depending on the relative conditional probabilities. Notice that this classification rule only requires relative expression between the two genes, which will make such biomarkers more robust and easy to interpret. Various studies have reported the success of TSP as a two-gene classifier [49–51]. However, for data sets with a complex phenotype, a single TSP pair may not be sufficient. The so-called k-TSP method has been proposed to make use of top k scoring pairs [52, 53]. Although a majority vote can be used to obtain a final classification, other supervised machine learning methods have been used and benchmarked including support vector machine, decision trees, naive Bayes classifier, k-nearest neighbor (k-NN), and prediction analysis of microarray (PAM) [52, 53].

A major approach to narrow down selected biomarkers is through variable selection in the framework of linear regression. The traditional stepwise variable selection approach only works for data with a small set of features and becomes computationally infeasible for big data such as microarray or RNAseq. Shrinkage estimators such as lasso (least absolute shrinkage and selection operator) have been developed to efficiently deal with such high-dimensional genomic data. Later efforts have extended the original lasso method including the grouped lasso by Yuan et al. where variables are selected or excluded in groups [54], the elastic net by Zou et al. which deals with correlated variables through a hybrid penalty [55], the graphical lasso by Friedman et al. for space covariance estimation [56], and the regularization paths for support vector machine [57].

Here we summarize the algorithms for lasso and closely related methods including elastic net and ridge regression. Given the response variable $Y \in R$ and a predictor vector $X \in R^p$ in a p dimensional space, we can approximate a linear function through $E\left(Y \middle| X = x\right) = \beta_0 + x^T\beta$ after observing N observation pairs (x_i, y_i) for $i = 1, 2, \ldots, N$ by solving the following optimization problem [58]:

$$\min_{(\beta_0,\beta)\in R^{p+1}} R_\lambda(\beta_0, \beta) = \min_{(\beta_0,\beta)\in R^{p+1}} \left[\frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 + x_i{}^T\beta\right)^2 + \lambda P_\alpha(\beta)\right]$$

where $P_\alpha(\beta)$ is the elastic net penalty term [55] defined as

$$P_\alpha(\beta) = (1-\alpha)\,\tfrac{1}{2}||\beta||_{\iota_2}^2 + \alpha||\beta||_{\iota_1}^1.$$

Here λ is a tuning parameter that users can specify or can be automatically calculated using cross validation based on prediction error.

Both the lasso algorithm (when $\alpha = 1$) and ridge regression (when $\alpha = 0$) are special cases of the elastic net method. Lasso provides coefficient estimates as either zero (for excluded variables) or nonzero (for selected variables) which is quite appealing for big genomic data. Ridge regression, on the other hand, only shrinks the coefficients and provides nonzero estimates only. For correlated variables, lasso tends to just pick one while ignore others. On the other hand, ridge regression allows borrowing information across all variables but retains all variables in the model. The elastic net with $0 < \alpha < 1$ enjoys the nice properties of both and usually performs better in genomic data. Elastic net has been efficiently implemented using cyclical coordinate descent and is publicly available in the R package glmnet.

Traditional strategies for biomarker discovery have focused on individual genes. However, tumorigenesis is a multi-step process involving sequential acquisitions of multiple genomic alterations regulated by different pathways and regulatory networks [62]. It is therefore appealing to identify biomarkers as sets of genes. According to Huang et al., three types of gene set analysis tools are available [63]. The first type is called singular enrichment analysis (SEA) which takes a preselected gene list as input and iteratively computes statistical enrichment of annotated gene sets by comparing them to random gene sets. The second type is called modular enrichment analysis (MEA) which considers inter-relationships as well as redundancies among annotated gene sets. MEA extends enrichment analysis from gene-centric or term-centric analysis to module-centric analysis which is more biologically plausible. The third type is called gene set enrichment analysis (GSEA). Different from SEA or MEA which requires a filtered gene list as input, GSEA takes into account all genes available and thus avoids the need of arbitrary cutoff for gene filtering. Different tools have different advantages and limitations. Users need to choose a tool that best fits their needs by considering the underlying statistical model, gene set annotation source, programming requirement, and output format.

Algorithms to identify cancer biomarkers are not limited to deal with a single data type. Several methods have been developed to integrate information across

data types. For example, integIRTy (integration using item response theory) is able to identify altered genes from multiple assay types accounting for multiple mechanisms of alteration [64]. integIRTy applies a latent variable approach to adjust for heterogeneity among different assay types for accurate inference. RABIT (regression analysis with background integration) is able to integrate public transcription factor (TF) binding profiles with tumor-profiling datasets [65]. RABIT controls confounding effects from copy number alteration, DNA methylation, and TF somatic mutation to identify cancer-associated TFs using a regression framework. Another interesting method is PARADIGM (PAthway Recognition Algorithm using Data Integration on Genomic Models) that integrates different genomic information based on pathway activity [66]. PARADIGM uses a factor graph to represent NCI pathway information which makes it effective to model different types of genomic data and various regulatory relationships.

6.2 Unsupervised Methods

The aforementioned methods are supervised since they require an outcome variable. There are also unsupervised methods for cancer biomarker discovery. Motivations for these methods originate from the fact that certain perturbations in the genome such as focal copy number change, gene fusions and mutations may lead to marked over-expression of oncogenes in a subset of samples. Since these oncogene activation events do not necessarily occur across all samples, traditional analytical approaches based on mean expression will fail [59]. Therefore, several methods have been proposed for this situation. For example, cancer outlier profile analysis (COPA) was developed to discover oncogenic chromosomal aberrations from outlier profiles based on median and median absolute deviation of gene expression. COPA identified the fusion of ERG and ETV1 which led to marked over-expression in 57 % of prostate cancer patients [59]. Later, a method called PACK (profile analysis using clustering and kurtosis) showed improved result by using Bayesian information criterion (BIC) and kurtosis [60]. Tong et al. developed SIBER (systematic identification of bimodally expressed genes using RNAseq data) using mixture model [61]. SIBER compares favorably to other methods and enjoys nice properties such as robustness, increased statistical power, and invariance to transformation [61]. We briefly summarize the SIBER algorithm here. Suppose the expression of a gene in sample s is e_s, SIBER models the distribution of gene expression $Pr(\mathrm{e_s})$ using a two-component mixture model each with mean expression μ_1, μ_2 and a shared dispersion parameter ϕ as following:

$$Pr\left(\mathrm{e_s}\right) = \pi f\left(\mathrm{e_s}; \mu_1, \phi\right) + (1-\pi) f\left(\mathrm{e_s}; \mu_2, \phi\right)$$

where π is the proportion of samples coming from the first component with density function $f(\mathrm{e_s}; \mu_1, \phi)$. The density function frequently used to model RNAseq data can be negative Binomial, generalized Poisson or log-normal distribution. After

estimating the parameters $(\pi, \mu_1, \mu_2, \phi)$, SIBER computes a generalized bimodality index BI as following:

$$BI = \sqrt{\pi\,(1-\pi)}\frac{\left|\mu_1 - \mu_2\right|}{\sqrt{(1-\pi)\,\sigma_1^2 + \pi\sigma_2^2}}$$

where σ_1^2, σ_2^2 were the variance of the two components. Through extensive simulation and real data analysis, Tong et al. showed that SIBER was a robust and powerful method to identify biomarkers with switch-like expression pattern [61].

7 Concluding Remarks

With recent advances in genomic technologies, the accumulation of genomic data is far exceeding Moore's law leading to the genomic data deluge. This represents a clear opportunity as well as pressing challenge for computational scientists to wade through the huge amount of data for biological insights. To identify biomarkers for cancer therapeutics, we should be familiar with relevant data resources and equip ourselves with effective computational tools. Given the extreme challenges for genomic data, the future success of cancer genomic research requires a continuous refinement and expansion of software tools and algorithms for the management, analysis, integration, and interpretation of high-throughput data.

References

1. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* cell, 2000. **100**(1): p. 57–70.
2. Davies, H., et al., *Mutations of the BRAF gene in human cancer.* Nature, 2002. **417**(6892): p. 949–954.
3. Samuels, Y., et al., *High frequency of mutations of the PIK3CA gene in human cancers.* Science, 2004. **304**(5670): p. 554–554.
4. Lynch, T.J., et al., *Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib.* New England Journal of Medicine, 2004. **350**(21): p. 2129–2139.
5. Paez, J.G., et al., *EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy.* Science, 2004. **304**(5676): p. 1497–1500.
6. Pao, W., et al., *EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(36): p. 13306–13311.
7. Weiss, R. *NIH Launches Cancer Genome Project*. 2005; Available from: http://www.washingtonpost.com/wp-dyn/content/article/2005/12/13/AR2005121301667.html.
8. Hudson, T.J., et al., *International network of cancer genome projects.* Nature, 2010. **464**(7291): p. 993–998.

9. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603–607.
10. Rees, M.G., et al., *Correlating chemical sensitivity and basal gene expression reveals mechanism of action.* Nature chemical biology, 2015.
11. Shoemaker, R.H., *The NCI60 human tumour cell line anticancer drug screen.* Nature Reviews Cancer, 2006. **6**(10): p. 813–823.
12. Yang, W., et al., *Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.* Nucleic acids research, 2013. **41**(D1): p. D955–D961.
13. Ding, L., et al., *Expanding the computational toolbox for mining cancer genomes.* Nature Reviews Genetics, 2014. **15**(8): p. 556–570.
14. Colburn, W., et al., *Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Biomarkers Definitions Working Group.* Clinical Pharmacol & Therapeutics, 2001. **69**: p. 89–95.
15. Frank, R. and R. Hargreaves, *Clinical biomarkers in drug discovery and development.* Nature Reviews Drug Discovery, 2003. **2**(7): p. 566–580.
16. Liang, M.H., et al., *Methodologic issues in the validation of putative biomarkers and surrogate endpoints in treatment evaluation for systemic lupus erythematosus.* Endocrine, metabolic & immune disorders drug targets, 2009. **9**(1): p. 108.
17. Leary, R.J., et al., *Development of personalized tumor biomarkers using massively parallel sequencing.* Science translational medicine, 2010. **2**(20): p. 20ra14–20ra14.
18. Ji, Y., et al., *Glycine and a Glycine Dehydrogenase (GLDC) SNP as Citalopram/Escitalopram Response Biomarkers in Depression: Pharmacometabolomics-Informed Pharmacogenomics.* Clinical Pharmacology & Therapeutics, 2011. **89**(1): p. 97–104.
19. CHEN, H.Y., et al., *Biomarkers and transcriptome profiling of lung cancer.* Respirology, 2012. **17**(4): p. 620–626.
20. Zhao, L., et al., *Identification of candidate biomarkers of therapeutic response to docetaxel by proteomic profiling.* Cancer research, 2009. **69**(19): p. 7696–7703.
21. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature Reviews Genetics, 2009. **10**(1): p. 57–63.
22. Pritchard, C.C., H.H. Cheng, and M. Tewari, *MicroRNA profiling: approaches and considerations.* Nature Reviews Genetics, 2012. **13**(5): p. 358–369.
23. Wright, P., et al., *A review of current proteomics technologies with a survey on their widespread use in reproductive biology investigations.* Theriogenology, 2012. **77**(4): p. 738–765. e52.
24. Mueller, C., L.A. Liotta, and V. Espina, *Reverse phase protein microarrays advance to use in clinical trials.* Molecular oncology, 2010. **4**(6): p. 461–481.
25. Strahl, B.D. and C.D. Allis, *The language of covalent histone modifications.* Nature, 2000. **403**(6765): p. 41–45.
26. Lund, A.H. and M. van Lohuizen, *Epigenetics and cancer.* Genes & development, 2004. **18**(19): p. 2315–2335.
27. Zuo, T., et al., *Methods in DNA methylation profiling.* Epigenomics, 2009. **1**(2): p. 331–345.
28. Soon, W.W., M. Hariharan, and M.P. Snyder, *High-throughput sequencing for biology and medicine.* Molecular systems biology, 2013. **9**(1): p. 640.
29. Barrett, T., et al., *NCBI GEO: mining tens of millions of expression profiles—database and tools update.* Nucleic acids research, 2007. **35**(suppl 1): p. D760–D765.
30. Barrett, T. and R. Edgar, *Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis.* Methods in enzymology, 2006. **411**: p. 352–369.
31. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets—update.* Nucleic acids research, 2013. **41**(D1): p. D991–D995.
32. Wilhite, S.E. and T. Barrett, *Strategies to explore functional genomics data sets in NCBI's GEO database*, in *Next Generation Microarray Bioinformatics*. 2012, Springer. p. 41–53.
33. Davis, S. and P.S. Meltzer, *GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor.* Bioinformatics, 2007. **23**(14): p. 1846–1847.
34. Kauffmann, A., et al., *Importing arrayexpress datasets into r/bioconductor.* Bioinformatics, 2009. **25**(16): p. 2092–2094.

35. Wu, L., et al., *Multidrug-resistant phenotype of disease-oriented panels of human tumor cell lines used for anticancer drug screening.* Cancer research, 1992. **52**(11): p. 3029–3034.
36. Garnett, M.J., et al., *Systematic identification of genomic markers of drug sensitivity in cancer cells.* Nature, 2012. **483**(7391): p. 570–575.
37. Cowley, G.S., et al., *Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies.* Scientific data, 2014. **1**.
38. Team, R.C., *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012*. 2014, ISBN 3-900051-07-0.
39. Huber, W., et al., *Orchestrating high-throughput genomic analysis with Bioconductor.* Nature methods, 2015. **12**(2): p. 115–121.
40. Durinck, S., et al., *Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.* Nature protocols, 2009. **4**(8): p. 1184–1191.
41. Durinck, S., et al., *BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.* Bioinformatics, 2005. **21**(16): p. 3439–3440.
42. Goecks, J., A. Nekrutenko, and J. Taylor, *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* Genome Biol, 2010. **11**(8): p. R86.
43. Blankenberg, D., et al., *Galaxy: a web-based genome analysis tool for experimentalists.* Current protocols in molecular biology, 2010: p. 19.10. 1–19.10. 21.
44. Reich, M., et al., *GenePattern 2.0.* Nature genetics, 2006. **38**(5): p. 500–501.
45. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Science signaling, 2013. **6**(269): p. pl1.
46. Rhodes, D.R., et al., *Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles.* Neoplasia, 2007. **9**(2): p. 166-180.
47. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proceedings of the National Academy of Sciences, 2001. **98**(9): p. 5116–5121.
48. Geman, D., et al., *Classifying gene expression profiles from pairwise mRNA comparisons.* Statistical applications in genetics and molecular biology, 2004. **3**(1): p. 1–19.
49. Youssef, Y.M., et al., *Accurate molecular classification of kidney cancer subtypes using microRNA signature.* European urology, 2011. **59**(5): p. 721–730.
50. Price, N.D., et al., *Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas.* Proceedings of the National Academy of Sciences, 2007. **104**(9): p. 3414–3419.
51. Xu, L., et al., *Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data.* Bioinformatics, 2005. **21**(20): p. 3905–3911.
52. Shi, P., et al., *Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction.* Bmc Bioinformatics, 2011. **12**(1): p. 375.
53. Tan, A.C., et al., *Simple decision rules for classifying human cancers from gene expression profiles.* Bioinformatics, 2005. **21**(20): p. 3896–3904.
54. Yuan, M. and Y. Lin, *Model selection and estimation in regression with grouped variables.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006. **68**(1): p. 49–67.
55. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005. **67**(2): p. 301–320.
56. Friedman, J., T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso.* Biostatistics, 2008. **9**(3): p. 432–441.
57. Hastie, T., et al., *The entire regularization path for the support vector machine.* The Journal of Machine Learning Research, 2004. **5**: p. 1391–1415.
58. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent.* Journal of statistical software, 2010. **33**(1): p. 1.
59. Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science, 2005. **310**(5748): p. 644–648.

60. Teschendorff, A.E., et al., *PACK: Profile Analysis using Clustering and Kurtosis to find molecular classifiers in cancer.* Bioinformatics, 2006. **22**(18): p. 2269–2275.
61. Tong, P., et al., *SIBER: systematic identification of bimodally expressed genes using RNAseq data.* Bioinformatics, 2013. **29**(5): p. 605–613.
62. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* cell, 2011. **144**(5): p. 646–674.
63. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic acids research, 2009. **37**(1): p. 1–13.
64. Tong, P. and K.R. Coombes, *integIRTy: a method to identify genes altered in cancer by accounting for multiple mechanisms of regulation using item response theory.* Bioinformatics, 2012. **28**(22): p. 2861–2869.
65. Jiang, P., et al., *Inference of transcriptional regulation in cancers.* Proceedings of the National Academy of Sciences, 2015. **112**(25): p. 7731–7736.
66. Vaske, C.J., et al., *Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.* Bioinformatics, 2010. **26**(12): p. i237–i245.