Chapter 2
Omic Data, Information Derivable and Computational Needs

Cancer is probably the most complex class of human diseases. Its complexity lies in: (1) its rapidly evolving population of cells that drift away from their normal functional states at the molecular, epigenomic and genomic levels, (2) its growth and expansion to encroach and replace normal tissue cells; and (3) its abilities to resist both endogenous and exogenous measures for stopping or slowing down its growth. According to Hanahan and Weinberg, cancer cells, regardless of the type, tend to have eight hallmark characteristics (Hanahan and Weinberg 2011). As introduced in Chap. 1, these hallmarks are: (1) reprogrammed energy metabolism, (2) sustained cell-growth signaling, (3) evading growth suppressors, (4) resisting cell death, (5) enabling replicative immortality, (6) inducing angiogenesis, (7) avoiding immune destruction, and (8) activating cell invasion and metastasis. Other authors have suggested some additional hallmarks of cancer such as tumor-promoting inflammation (Colotta et al. 2009) and deregulated extracellular matrix dynamics (Lu et al. 2012). These recognized hallmarks have provided an effective framework for addressing cancer-related questions, having led to a deeper understanding of this disease. However, the reality is that our overall ability in curing cancer has not yet made substantive improvements, particularly in adult cancers that account for 99 % of all cancers since the start of the "War on Cancer" in 1971 (The-National-Cancer-Act 1971).

Major challenging issues that clinical oncologists have to deal with include not only considerable heterogeneity and different genetic backgrounds even within the same type of cancer, but also that effective medicines tend to lose their efficaciousness within a year, or often within a few months. A natural question to pose is: *What are the reasons for this loss of effectiveness*? Intuitively this is due to a cancer's ability to evolve rapidly, particularly in terms of generating drug-resistant sub-populations, which is facilitated by its abilities to proliferate and to accumulate genomic mutations rapidly. However, such an answer, plausible as it may be, has possibly missed the real root issue: *Why do these cells divide so rapidly in the first place*?

Y. Xu et al., *Cancer Bioinformatics*, DOI 10.1007/978-1-4939-1381-7_2

The *Red Queen Hypothesis*, proposed by Leigh Van Valen in 1973, may provide a good framework for studying this and other cancer-related fundamental issues from an evolutionary perspective. The hypothesis states: an adaptation in a population of one species may change the selection pressure on a population of another species, giving rise to an antagonistic coevolution (Valen 1973). When in this frame of thinking, one may be inspired to ask: *What specific selection pressures must the evolving neoplastic cells overcome, pressures that may drive their rapid proliferation*? Currently we do not have an answer to this question yet. Among the many reasons that our knowledge is so sparse has been the lack of molecular-level data, full analyses and mining of which can potentially reveal the full complexity of an evolving cancer. While large quantities of *omic* data such as gen*omic*, epigen*omic*, transcript*omic*, metabol*omic* and prote*omic* data have been generated for a variety of cancer types, only a few cancer studies have been designed to take full advantage of all the information derivable from the available *omic* data (Cancer-Genome-Atlas-Research 2008, 2011, 2012a, b, c, 2013a, b; Kandoth et al. 2013). Integrative analyses of multiple data types may prove to be essential to gain a full and systems-level understanding about a cancer's evolution dynamics, including the elucidation of its true drivers as well as key facilitators at different developmental stages of a cancer. We anticipate that only when all of the key information hidden in *omic* data can be fully derived and utilized can we expect a meaningful breakthrough in our understanding of cancer.

2.1 Genomic Sequence Data

The Human Genome Project was initiated in 1986 by the US Department of Energy and the National Institutes of Health, which ultimately led to the generation of the first digital copies of two complete human genomes in 2001 (Lander et al. 2001; Venter et al. 2001), one by government agencies and one by a private organization. For the first time in history, the three billion base pairs (bps) of nucleotides comprising a complete human genome are represented in a digital form, directly readable by humans and computers, allowing researchers and clinicians to view and analyze the detailed genetic makeup of two healthy humans. This singular achievement has profoundly changed biological and medical sciences, clearly representing the most significant discovery since the finding of the double-strand helical structure of DNA in 1950s. Complementing and extending the invaluable genome sequence data, the major change that the Human Genome Project has brought about is that genetic science is now equipped with two powerful tools: rapid genome-sequence generation and computation-based information discovery from the genomic sequences. These tools along with the advances they have helped to make in the biological sciences, have fundamentally transformed the science of genetics, which is now data-rich and quantitative. This transition has attracted and continues to attract many mathematical and computational scientists to study problems related to genomes and other biomolecules represented in digital forms. The progress made has further transformed

the general biological sciences and has substantially advanced our overall ability to study more complex biological problems than could be done before the *omic* era.

With the public availability of digitally represented human genomes in hand, scientists have computationally identified the vast majority of the ~20,000 protein-encoding genes in our genome, along with large numbers of single-nucleotide polymorphisms (SNPs) and other types of genetic variations across individuals and different ethnic groups as well as various disease groups. Targeted sequencing of specific genomic regions deemed to be relevant to certain diseases has led to the identification of numerous genetic markers for multiple diseases. For example, Down syndrome is now understood to be caused by an extra copy of chromosome 21. A few additional examples include: (1) adrenoleukodystrophy, a progressive degenerative myelin disorder caused by mutations in the *ABCD1* (ATP-binding cascade subfamily D) gene, which was made popular because of the movie "Lorenzo's Oil" in the early 1990s; (2) a class of hereditary breast cancers caused by mutations in the *BRCA* (breast cancer) genes; (3) familial hyperlipidemia attributable to mutations in the *APC* (adenomatous polyposis coli) gene; and (4) frontotemporal dementia, a form of inherited dementia, caused by mutations related to the splicing of exon 10 of the *Tau* gene (D'Souza et al. 1999). All these were detected through genome-scale or targeted gene sequencing and associated sequence analyses.

In addition to the Human Genome Project, a number of closely related genome sequencing projects have been established to provide a more comprehensive dataset for the human genome(s): (1) the Human Genome Diversity Project to document genomic differences across different ethnic groups (Cavalli-Sforza 2005); (2) the Human Variome Project to establish relationships between human genomic variations and diseases (Cotton et al. 2008); (3) the International HapMap Project to develop a haplotype map of the human genome (International-HapMap 2003); (4) the 1000 Genome Project to establish a detailed catalog of all human genetic variations (Service 2006); and (5) the Personal Genome Project to sequence the complete genomes and establish the matching medical records of 100,000 individuals (Church 2005). All these sequencing projects, along with other related ones, such as the Neanderthal Genome Project (Green et al. 2010) and the Chimpanzee Genome Project (Cheng et al. 2005; Green et al. 2010), could provide a comprehensive view of the genomes of healthy humans with normal polymorphisms as well as mutations associated with various diseases.

The Cancer Genome Atlas (TCGA) represents probably the most ambitious cancer-genome sequencing project, which aims to sequence up to 10,000 cancer genomes covering 25 major cancer types by 2014 and make the data publicly available (Cancer-Genome-Atlas-Research et al. 2013). Such data will provide a substantial amount of information about cancer-related genomic mutations. By comparing the genome sequences of a cancer and the matching normal tissue, one can identify all the genomic changes in the cancer genome, which generally fall into two categories: simple and complex mutations. Specifically, *simple* mutations refer to single base-pair mutations and DNA single or double-strand breaks; and *complex* mutations refer to duplications and deletions (together referred to as *copy-number changes*), translocations and inversions of genomic segments. Simple mutations can

result from exogenous factors such as radiation, air-borne and food-related carcinogens in the environment, as well as from endogenous factors in the microenvironments inside our bodies, including ROS (reactive oxygen species) and other reactive metabolites plus random mutations. For example, ionizing radiation, including X-rays and gamma rays, can directly cause point mutations and DNA breaks. In addition, a variety of non-radioactive carcinogens have been identified that can damage DNA, including microbes, chemical compounds in the environment and reactive species inside our cells, as detailed in Chap. 5. Free radicals represent a large class of internal, potentially carcinogenic agents that are highly reactive molecules and can participate in undesired reactions, causing damages to cells and specifically to DNA. Infidelity of transcription and/or repair can also lead to simple mutations. While these carcinogens can produce simple DNA damages, it is the faulty or imprecise DNA replication and repair machineries that lead to the complex mutations, namely undesired duplications, deletions, inversions and translocations of large DNA segments.

There are multiple situations that can result in such complex genomic mutations. For example, under persistent hypoxic conditions, cells tend to use emergency mechanisms to repair simple mutations, but the inaccuracy of such mechanisms can lead to complex mutations as defined above (Scanlon and Glazer 2013). Here we outline one such mechanism, named *microhomology-mediated end joining* (MMEJ) for repairing double-strand DNA breaks, through which undesired DNA copy-number changes, inversions and translocations can result (Truong et al. 2013). Like the regular repair mechanism for double-strand breaks, MMEJ uses the sister chromosome as the template to replace the region with a break. The difference is that it uses a much shorter homologous region in the sister chromosome, typically 5–25 bps rather than the usual 200 bps required by the normal DNA repair mechanism, hence the designation microhomology-mediated. While the advantage is that this mechanism is substantially faster than the regular DNA repair machinery, which is needed under certain emergency conditions, it is error prone due to the less stringent requirement for finding the equivalent region in the sister chromosome, thus leading to various complex mutations (Bentley et al. 2004). This mechanism is used only under highly stressful conditions when the regular DNA repair mechanisms are functionally repressed (Bindra et al. 2007), and hence is often used in cancer-associated environments.

Knowing how different genomic mutations occur, one could possibly develop computational models to infer the evolutionary history of the mutations observed in a cancer genome from the matching reference genome. The idea is that one can first identify all the genomic differences between a cancer genome and the matching reference genome. For each identified complex mutation, one can apply a mechanistic model like the one outlined above (or from the literature) to predict how it occurs from the previous generation of the genome, while simple mutations can be assumed to take place randomly according to some stochastic models. It is worth noting that some of the evolutionary intermediates (mutations) may or may not be present in the cancer genome, due to the possibilities that some portions of the genome might have been deleted during evolution. In addition, it should be emphasized that such an

approach (even when taking into consideration the other emergency DNA repair mechanisms) may not necessarily yield a unique evolutionary path from the reference to the cancer genome. One possible way to constrain this phylogenetic reconstruction problem to a solution space as small as possible is to find such a path under the parsimony assumption (Steel and Penny 2000), as often used in phylogenetic reconstruction algorithms. Specifically one can require that the predicted evolutionary path have either the smallest number of generations or the highest consistency with the occurrence probabilities of different types of mutations as documented in the literature. As of now, no such algorithms have been published for making evolutionary path predictions, but the need for such tools is clearly there in order to understand the evolution of a cancer genome.

Various other types of information may also be derivable from cancer genomes, such as: (1) oncogenes and tumor suppressor genes (see Chap. 1 for definition) that may be specific to a particular cancer type. Examples include gene fusions as in the case of the Philadelphia chromosome for chronic myelogenous leukemia (CML) (Nowell and Hungerford 1960); (2) potential integration of microbial genes into the cancer genomes as in the case of hepatitis B virus genes integrated into the host genome; (3) biological pathways that are enriched with genetic mutations in a particular cancer, leading to the loss of function at the pathway level; and (4) changes in mutation patterns as a cancer advances.

By systematically identifying mutations in the genomes of multiple patients of the same cancer type, one can identify biological pathways enriched with such mutations, using analysis tools like DAVID (Huang et al. 2009) against pathway databases such as KEGG (Kanehisa et al. 2010, 2012, 2014), BIOCARTA (Nishimura 2001) or cancer-related gene sets (Forbes et al. 2011; Chen et al. 2013; Zhao et al. 2013). For example, a study, published in 2007 on genomic mutations observed across 210 cancer types, discovered that the pathway having the highest enrichment with non-synonymous mutations is the *FGF* (fibroblast growth factor) signaling pathway, revealing one commonality among changes needed by cancer evolution across different cancer types (Greenman et al. 2007). With such information, one can further infer which cellular processes need to be terminated or become hyperactive in any specific order as a cancer evolves, hence possibly developing new insights about the evolutionary paths unique to particular cancer types or common among all cancer types.

2.2 Epigenomic Data

Epigenomic data provide information about all the chemical modifications on the genomic DNA and associated histone proteins in a cell, namely *DNA methylation* and *histone modification*, among a few other less studied epigenomic activity types. While epigenetic analyses are not new, it is the high-throughput array and sequencing techniques that have made such analyses at a genome scale possible and have clearly advanced our overall capabilities to study cancer.

DNA methylation is a process by which a methyl group is added to the carbon 5 position of cytosine residues (C) in CpG dinucleotides. This is accomplished through a group of enzymes known as *DNA methyl-transferases*, the reactions of which can be reversed by another group of enzymes termed *DNA demethylases*. When a CpG region is highly methylated, they attract a group of enzymes called *histone deacetylases* that will initiate chromatin remodeling to change the local structure of the DNA, hence changing its accessibility to large molecular structures such as the transcription machinery, RNA polymerase. Since long CpG regions (denoted as *CpG islands*) tend to be associated with the promoters of genes, methylation of such regions represses the expression of the genes.

Histones are proteins that bind with DNA to form the basic folding units, denoted as *nucleosomes*, of chromatin, as introduced in Chap. 1. The packing density of chromatin is closely related to the transcriptional state of a gene, i.e., lower packing density implying higher transcriptional activity. Cells change their chromatin structures through post-translational modifications on the relevant histones, including acetylation, deamination, methylation, phosphorylation, SUMOylation and ubiquitination. The understanding is that interactions between histones and DNA are formed by electrostatic attraction between the positive charges on the histone surface and the negative charges on DNA. Consequently, modifications on histones may change the charges of the surface residues, possibly changing the conformation and the transcriptional accessibility of a folded DNA and ultimately enhancing or repressing expression of the relevant genes (Strahl and Allis 2000; Kamakaka and Biggins 2005). Another mechanism is through recruiting and applying chromatin remodeling *ATPases*, where histone modifications can lead to disruptions of *ATPase* attraction to the chromatin, hence altering the DNA's physical accessibility to the RNA polymerase (Vignali et al. 2000).

Various techniques have been developed to reliably capture DNA methylations and histone modifications at a genome scale. Among the assays that have been used for detecting methylations is the *bisulfite* sequencing technique (Yang et al. 2004). By converting each methylated C to a T and removing the methylation, the bisulfite method utilizes the current sequencing techniques to produce the modified sequence and then recovers the methylation locations through comparisons between the sequenced Ts and Cs at the same locations in the original DNA and the modified DNA done as above.

Histone modification sites can be detected using the ChIP-chip array technique (Huebert et al. 2006), which has previously been used to identify the binding sites of transcriptional factors. The difference here is to detect the DNA binding sites with histones relevant to the packing of DNA. Comparisons between the identified DNA binding sites under different conditions can lead to the identification of modified chromatin structures. The advancement of sequencing techniques in the past few years has led to the development of the second generation ChIP technique, namely *ChIP-seq*, which can provide more quantitative and reliable data about histone modification sites.

From either of the two types of epigenomic data, one can infer genes that are primed to be repressed or enhanced transcriptionally at the epigenomic level.

These data, in conjunction with other *omic* data such as transcriptomic and genomic information, can be used to derive association relationships between epigenomic activities and the cellular as well as micro-environmental states. This can lead to identification of possible triggers and regulatory pathways of different epigenomic activities. Information of this type is clearly needed since, although numerous epigenomic effectors such the enzymes for DNA methylation and histone modifications have been identified, very little is known about the regulation of these effectors and under what conditions a specific set of genes will be methylated. As discussed in Chap. 9, epigenomic level changes can be considered as an intermediate step between (reversible) functional state changes of effector molecules and (permanent) genomic mutations. A detailed discussion regarding the possible relationships among these three types of changes needed by evolving cancer cells is given in Chap. 9.

A number of large-scale epigenomic sequencing projects have been initiated with similar ambitious goals to those of the genome sequencing projects outlined in Sect. 2.1. These projects include: (a) the NIH Roadmap Epigenomics Program, which started in 2008 with the aim of producing histone modification data for over 30 types of modifications in a variety of human cell types; (b) a component of the ENCODE (Encyclopedia of DNA Elements) project launched by the US National Human Genome Research Institute aiming as part of its goal the characterization of the epigenomic profiles of 50 different tissue types; (c) the International Human Epigenome Consortium having its goal to build on and expand the NIH Epigenomics Program to include nonhuman cells and tissues, and to make it a functional international program; and (d) some regional epigenomics projects such as the "Epigenetics, Environment and Health" project in Canada and the Australian Alliance for Epigenetics. A number of human epigenomic databases have been developed as the result of these and related projects (see Chap. 13 for details).

2.3 Transcriptomic Data

The advent of microarray technology in the mid-1990s has made it possible to measure in real time the expression levels of all the genes encoded in the human genome under defined cellular conditions. This methodology also applies to other genomes as long as their protein-encoding genes are known. This is one of the high-throughput techniques that has clearly fueled the revolution in biological sciences that we have been witnessing since the start of the Human Genome Project.

Comparative analyses of gene-expression data of cells collected under different controlled conditions or on disease *versus* control tissues can provide a large amount of information useful for studying human diseases at the molecular and the cellular levels. For example, by comparing gene-expression levels in a lung cancer tissue with those in the adjacent healthy tissue of the same patient, one can identify differentially expressed genes in the lung cancer *versus* the healthy lung. While not necessarily all the differentially expressed genes are directly relevant to cancer, this

information provides a basis for further inference of genes that may be directly relevant to cancer. For example, one can compare such sets of differentially expressed genes across multiple patients of the same cancer type to eliminate those genes whose differential expressions are specific to a few individuals or cancers at a specific developmental stage. That is, one can identify genes that may be most essential to the development of a cancer type through the identification of genes that are commonly differentially expressed across all or the majority of the patients of the cancer type examined.

When applied in conjunction with pathway-enrichment analysis, particularly against the eight cancer-hallmark related pathways mentioned earlier, one can identify hallmark pathways enriched with up-regulated (or down-regulated) genes in a specific type of cancer. If the cancer data also have the stage information, one can further derive information about how each of the cancer hallmarks is executed at the molecular and cellular levels for this cancer type and in what order. By comparing such information across multiple cancer types, one can possibly detect which relative orders among the observed hallmark events are essential and which are coincidental. And by comparing such data between two subgroups of patients of the same cancer type, for example one with smoking histories and the other without, one can possibly derive how smoking may have contributed to the development of individual hallmark events. Similar analyses can be used to discover possible contributions by other lifestyle habits.

Actually, much more information can be derived through analyses of cancer transcriptomic data. For example, tiling array is a variation of the gene-expression technique used to detect DNA-binding sites of specific proteins through ChIP-chip experiments, hence making it possible to identify transcription regulators of specific genes under particular conditions (Ren et al. 2000; Iyer et al. 2001). RNA-seq is the new generation of techniques for transcriptomic data collection (Wang et al. 2009). It refers to the use of high-throughput technologies to sequence cDNAs that are reversely transcribed from the expressed RNA molecules. By doing deep sequencing, the dynamic range of RNA-seq can span five orders of magnitude, substantially larger than those of microarray-based techniques. This allows more accurate identification of differentially expressed genes, particularly those that tend to express at a relatively low or high level and where changes tend to be relatively small but statistically significant, such as those often observed with transcription factors. In addition, RNA-seq techniques are digital in nature, in comparison with the analog signals provided by microarrays. One advantage of digital signals is that the resulting measurements are more repeatable compared to analog signals and less prone to be affected by the experimental environments. The biggest advantage of RNA-seq data over microarray data is that it contains all the information about alternatively spliced variants since they do not rely on short sequence probes as in microarrays, instead producing the entire sequence for each transcript. Such information allows one to derive all splicing variants in specific cancers and cancer stages, thus enabling more detailed functional mechanism studies.

A few computer programs have been developed and made publicly available for inference of splicing variants based on RNA-seq data, such as Cufflinks, which requires a reliable reference genome for its inference of splicing isoforms (Trapnell

et al. 2010). Another popular transcript-assembly program, Trinity, is a *de novo* method, i.e., no reference genome is required (Grabherr et al. 2011), but at the expense of less reliable assembly results compared to Cufflinks. The limitation of Cufflinks and similar programs is that they may not necessarily work well on cancer RNA-seq data when the underlying cancer genome is not available, which could be substantially different from the matching genome of healthy cells since cancer genomes tend to have large numbers of genomic reorganizations such as translocations, copy-number changes and inversions, as well as breaks as discussed in Sect. 2.2. Thus, more effective computational techniques are clearly needed for inference of splicing isoforms from cancer RNA-seq data.

Presently, a number of databases for microarray and RNA-seq gene-expression data have been developed and are publicly available. For example, GEO is a general-purpose gene-expression database consisting of both cancer and other tissue types (Edgar et al. 2002). A cancer-centric genome database that also contains epigenomic and transcriptomic data for numerous cancer types is TCGA (Cancer-Genome-Atlas-Research et al. 2013). Gene Expression for Pediatric Cancer Genome Project is a gene-expression database developed specifically for pediatric cancers (Downing et al. 2012). Overall these databases have genome-scale transcriptomic data for over 200 different types of cancer tissues and a substantially larger number of cancer cell lines. A tremendous amount of information could potentially be derived through comparative analyses of these data across different cancer types and cancers at varying stages or of distinct malignancy grades (see Chap. 3). For example, by simply plotting the average number of differentially expressed genes across cancer samples *versus* the 5-year survival rate for each of the following nine cancer types: melanoma, pancreatic, lung, stomach, colon, kidney, breast, prostate cancers and basal cell carcinoma, one can see that there is a close relationship between these two numbers (see Fig. 2.1).

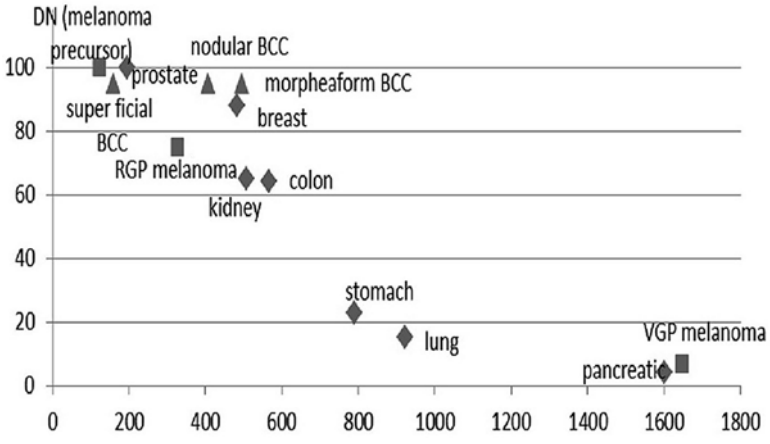

Fig. 2.1 The 5-year (y-axis) survival rate for each cancer type *versus* the average number of differentially expressed genes per cancer sample (x-axis) (adapted from (Xu et al. 2012))

By examining the average up- or down-regulation levels of genes in selected pathways of different cancer types, it is possible to derive information about activated energy metabolism in different cancer types, which vary from glucose- to lipid- to amino acid-based. As an example, Fig. 2.2 shows the activity levels of multiple energy producing metabolic pathways, covering glycolysis, the TCA cycle, oxidative phosphorylation and fatty acid metabolism in nine cancer types. One can see from the figure that pancreatic cancer has the highest up-regulation in glucose metabolism, followed by kidney and lung with breast cancer having the least changes in glucose metabolism when compared with their matching control tissues. One can also see that, while most of the seven cancer types on the left show down-regulation or no changes in oxidative phosphorylation, both skin cancer types, namely melanoma and basal cell carcinoma, show up-regulation in this pathway. [*N.B. Throughout this book, all the analyses of transcriptomic data across different patients samples are properly normalized, hence comparisons among fold-changes of genes across different samples are meaningful.*]

A variety of computational techniques have been developed for information derivation from gene-expression data, including: (1) identification of differentially expressed genes using simple statistical tests such as T-test or Fisher's exact test, (2) clustering analysis, (3) bi-clustering analysis and (4) pathway enrichment analysis for differentially expressed genes. The following discussion provides some basic ideas about these analysis techniques, followed with a list of novel techniques for more advanced analysis needs.

2.3.1 Data Clustering

Identification of co-expressed genes is a basic technique for gene-expression analysis, which has a wide range of applications in cancer studies. The idea is to identify all genes whose expression patterns exhibit statistical correlations over a time course (typically for cell line-based data) or among a collection of samples; such genes are called *co-expressed genes*. There are numerous online tools for identification of co-expressed genes such as DAVID, CoExpress (Nazarov et al. 2010) and GeneXPress (Segal et al. 2004). Co-expressed genes may suggest that the genes are transcriptionally co-regulated even though some co-expressed genes appear coincidentally, particularly when the number of conditions or the number of samples is small. One way to computationally "validate" such a prediction is through identification of conserved *cis* regulatory motifs within the promoter sequences of the co-expressed genes (Liu et al. 2009). The rationale is that if these genes are indeed co-regulated transcriptionally, they should share conserved *cis* regulatory elements for binding with their common transcription regulators. From the predicted co-expressed genes and *cis* regulatory motifs, one can predict with confidence that these genes are transcriptionally co-regulated, and even possibly predict their main transcription regulators using tools such as those by Essaghir et al. (2010) or by Qian et al. (2003).

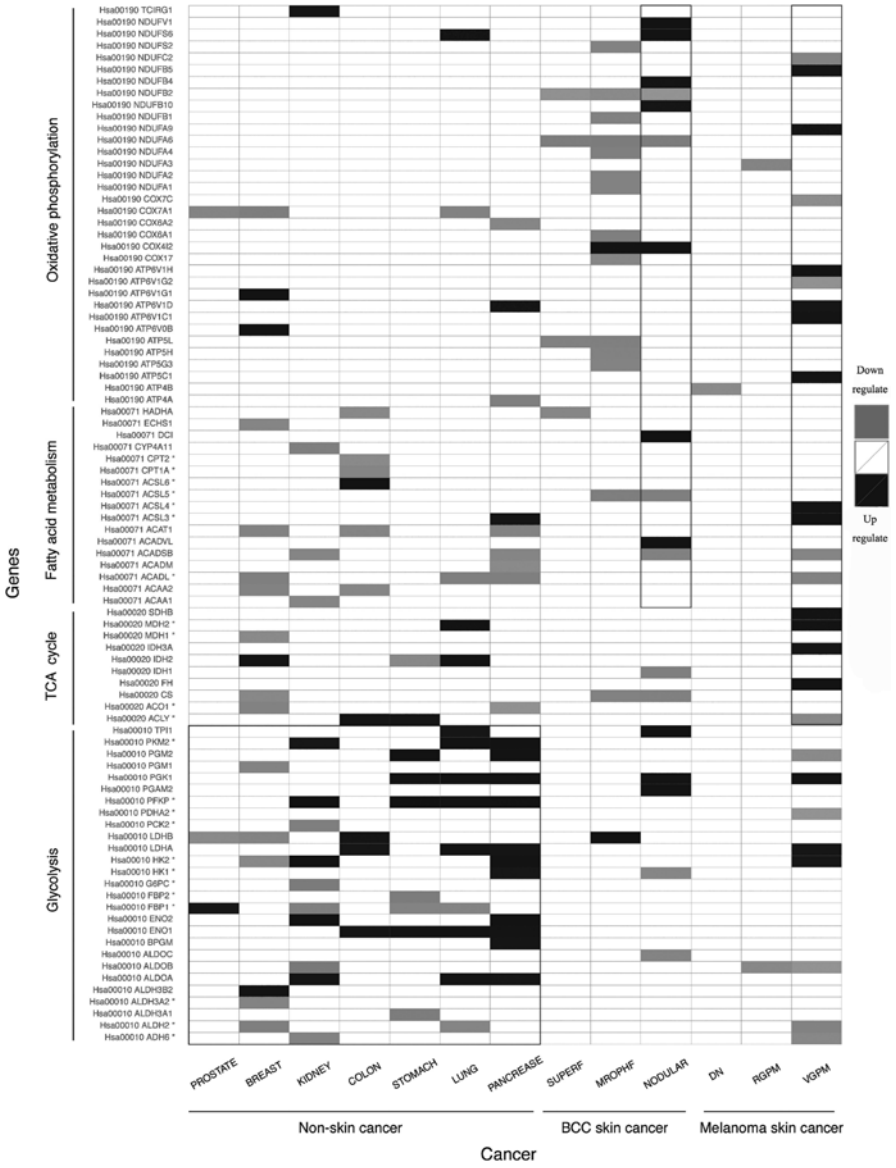

Fig. 2.2 Gene-expression levels associated with the energy metabolism of glucose (both glycolytic fermentation and oxidative phosphorylation) and fatty acids plus the TCA cycle in nine cancer types. The y-axis is a list of genes involved in four metabolic pathways: oxidative phosphorylation, fatty acid metabolism, TCA cycle and glycolysis; and the x-axis is a list of nine cancer types, including three stages of basal cell carcinoma (BCC) and melanoma, respectively. Each entry is the average log-ratio of expression levels between cancer samples and the matching control samples in different cancer types. The side-bar on the *right* shows the gray-level code for the expression level changes, with "*gray*" indicating down-regulation, "*white*" for no change and "*black*" denoting up-regulation. Adapted from (Xu et al. 2012)

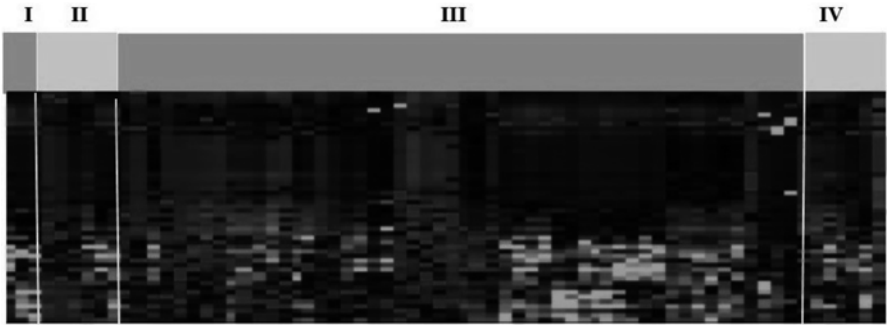

Fig. 2.3 A heat-map of gene-expression changes of 42 genes, with each row representing one gene and 80 gastric cancer samples *versus* the matching control samples, with each column representing one sample, which are grouped into four stages: I, II, III and IV, with *light gray*, *dark gray* and *black* representing up-, down-regulation and no changes, respectively. The figure is adapted from (Cui et al. 2011)

2.3.2 Bi-clustering Analysis

Bi-clustering is a generalized form of clustering analysis, which aims to identify co-expressed genes among some to-be-identified subgroups of samples, but not among all samples. Such a technique is particularly useful for discovering subtypes, stages or grades of a cancer (see Chap. 3 for details). Figure 2.3 shows one example of signature genes for gastric cancer stages identified through a bi-clustering analysis. Specifically, 42 genes are found to exhibit distinct patterns for a group of 80 gastric cancer samples (one sample from each patient) grouped according to their stages (Cui et al. 2011). Interestingly the samples assigned to stage III exhibit two distinct expression patterns, with samples on the left clearly showing different patterns from those on the right, suggesting that these patients may actually fall into five different stages such as stage I, II, IIa, III and IV, rather than four as proposed by the pathologists who analyzed these samples (Cui et al. 2011).

A bi-clustering problem is computationally much more difficult to solve than a clustering problem since it involves two variables, i.e., genes to be identified as co-expressed and samples to be found to have similar expression patterns, compared to only one variable, i.e., co-expressed genes in traditional clustering analyses. A few computer tools have been published for identification of bi-clusters in gene-expression datasets such as QUBIC (Li et al. 2009) and BicAT (Barkow et al. 2006). After bi-clusters are identified, similar analyses about regulatory relationships can also be carried out as above to predict the possible transcription regulators for each bi-cluster.

2.3.3 Pathway (or Gene Set) Enrichment Analysis

Pathway enrichment analysis is a way to map up- or down-regulated genes to higher level functional organizations such as biological pathways, networks or gene sets that are each known to be relevant to cancer or cancer-related. The basic idea is to homology-map the identified up-regulated (or down-regulated) genes to known pathways in pathway databases such as KEGG, REACTOME (Croft et al. 2011) or BIOCARTA, and then assess if a specific pathway has substantially more genes mapped to it than by chance, measured using statistical significance values. For example, DAVID is one popular tool for doing pathway enrichment analysis. Basically, it homology-maps a set of given genes to pathways in the above databases, then assesses the statistical significance of having K given genes in the given set mapped to a specific pathway using κ statistics, i.e., a chance-corrected measure of co-occurrence, and predicts that a pathway is enriched by the given gene set if its statistical significance is above some threshold (Huang et al. 2007). Figure 2.4 shows one enriched pathway by up-regulated genes in gastric cancer.

With the increasing needs for studying more complex analysis problems based on gene-expression data, there is clearly an urgent necessity for more powerful analysis techniques. A few are listed here, which could definitely benefit from the involvement of researchers equipped with advanced statistical analysis techniques.

1. *Inference of causal relationships*: Analyses discussed above, such as clustering or bi-clustering, can provide association relationships among activities of genes or pathways through detection of correlations among their expression patterns. Clearly cancer researchers could benefit even more if such analyses can be extended to infer causal relations among genes or pathways with altered expression patterns.

   Causality has been difficult to derive due to the nature of the problem (Pearl 2009). Many may remember the argument made by the tobacco industry when being presented with statistical data showing that smokers have higher probabilities of developing lung cancer than non-smokers. The industry officials argued that such data do not necessarily imply that smoking causes cancer, pointing out the following: there could be an unknown genetic factor that gives rise to a sub-population who enjoys smoking and has a higher propensity to develop lung cancer. Logically, this argument holds. Hence, in order to prove that smoking indeed causes lung cancer, one would need to demonstrate that individuals who are forced to smoke, regardless if they like it or not, are at higher risk of developing lung cancer than those who are forced not to smoke. This would then rule out a possible contribution from genetic factors as suggested by the defense lawyers of the tobacco industry. In general, inference of causality is fundamentally hard. Fortunately, there have been some interesting developments in theoretical studies on causal relationships. One example is the development of *causal calculus* by Pearl (2009). Application of this or other causal theories to the information-rich gene-expression-based causality analyses would help to advance the field in a profound way.

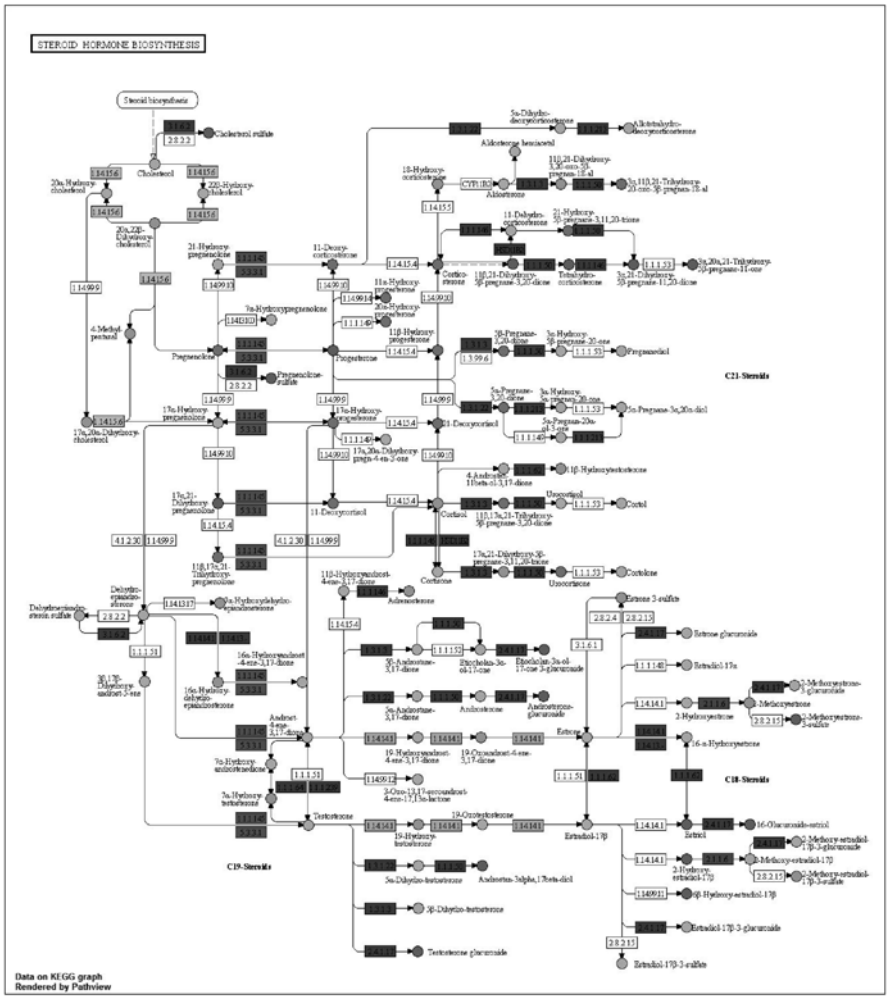

Fig. 2.4 An example of a KEGG pathway enriched with differentially expressed genes in gastric cancer *versus* matching controls. Each *rectangle* represents an enzyme-encoding gene and each *oval* represents a metabolite. An up-regulated gene is marked as *dark gray* and down-regulated gene marked in *light gray* while a *white rectangle* represents an enzyme whose gene is not identified yet. A metabolite with increased concentration is marked in *dark gray* and a decreased metabolite is marked in *light gray*

2. *De-convolution methods for expression data collected on cancer tissues*: One challenge in analyzing gene-expression data collected on cancer tissues is that the data are not from a homogeneous cell population, but instead a collection of different cell types with cancer cells as the dominating sub-population. It is well known that each sample of cancer tissue generally has other cell types such as macrophages and other immune cells, stromal cells, and blood vessel cells,

although there may have been attempts to make the cell population as homogeneous as possible, using techniques such as laser-directed microdissection (Emmert-Buck et al. 1996). The reality is that collecting highly homogeneous cell populations from cancer tissues is challenging and very time consuming.

Gene-expression data collected on a mixture of multiple cell types can easily lead to false conclusions if done without proper data processing. This issue has been reflected in a common complaint from gene-expression data analysts that tissue gene-expression data are not reliable and are difficult to compare across different samples. One key reason is that tissue samples collected by different labs may have been processed using different procedures so that the subpopulations of different cell types may be different from those *in situ.* Moreover, different sample-processing procedures may lead to systematic changes in the sub-populations but in different ways, thus making tissue gene-expression data not easily comparable.

It is our belief that techniques in statistical analysis, properly applied, can aid immensely in resolving the issue by de-convoluting the observed gene-expression data into expression levels contributed by different cell types. The basic idea of one such de-convolution technique is as follows. Each cell type has its unique functional characteristics. For example, cancer cells are the only cell type in the tissue that divides rapidly, while fibroblasts are the only cell type that synthesizes the components of the extracellular matrix. These unique functional characteristics of different cell types are reflected by their gene-expression data. Specifically, it is expected that each cell type can be represented (or approximated) by a set of expressed genes unique to the cell type, along with the cell type-specific correlations among the expression levels of different genes. Such cell type-specific (condition-invariant) correlations among their genes can possibly be represented in some generalized form of a covariance matrix, which can be considered as the *signature o*f individual cell types. To derive such a signature, one needs unambiguous gene-expression data of specific cell types collected under multiple and different conditions, allowing the capture of the invariance among the correlations between expression patterns of individual genes.

With such a reliable de-convolution tool, one can decompose each gene-expression dataset collected on cancer tissues into gene-expression contributions from different cell types. Then, one can analyze the gene-expression data predicted to be solely associated with cancer cells or other cell types such as macrophages to understand the interplay between cancer and immune cells. Such decomposed datasets of cancer samples at different stages have the potential of enabling one to realistically study a range of important problems in elucidating the complex relationships among different cell populations in each cancer niche, which are not feasible with the current experimental techniques.

3. *Development of an infrastructure in support of the study of cancer systems biology*: Another area where computational statistical techniques can make a fundamental contribution is in characterization of cancer microenvironments and in linking micro-environmental factors to the evolutionary trajectories of specific

cancer tissues. While experimental studies of the evolving microenvironment of a cancer *in vivo* may not be feasible, computational analyses of gene-expression data could help to solve such a problem. The premise is as follows. When the microenvironment changes, such as changes in (1) the composition and physical properties of the pericellular matrix, (2) the level of hypoxia, (3) the ROS level, (4) the pH level and (5) the sub-population sizes of different cell types in the stromal compartments (see Chap. 10), some genes will respond by changing their expression levels. For example, when the cellular level of oxygen changes, the expression patterns of the *HIF1* (hypoxia-induced factor) and *HIF2* genes change, as discussed in Chap. 1. By carefully analyzing gene-expression data collected under specific conditions on relevant cancer cell lines, one should be able to train predictors for changes in each aspect of the microenvironment based on their relationships reflected by gene-expression data. Such prediction capabilities will enable cancer researchers to examine how micro-environmental factors change as a cancer evolves and to link such information to cancer phenotypes, hence possibly generating new understanding about how microenvironments affect cancer progression and cancer phenotypes.

2.4 Metabolomic Data

Our own experience has been that transcriptomic data represent probably the most information-rich data that are relatively straightforward to obtain for cancer studies. Such data are particularly useful for gaining a big-picture view and for the derivation of rough models for a specific mechanism, while genomic and epigenomic data can provide useful complementary information. Transcriptomic data, however, do not always portray an accurate picture regarding the activity of a pathway. This is because they measure only the intermediates for making the functional parts, the proteins, of the pathway; others, such as those constitutively expressed, will of course not appear as altered gene expressions. Clearly, it is highly desirable to have protein expression data. However, proteins are notoriously difficult to study, much more complex than, say, transcripts, as proteins may have different post-translational modifications and splicing variants, which are not amenable to the current high-throughput techniques. Consequently, proteomic data have not been as widely used as transcriptomic data in cancer studies. Metabolomic data can, however, assist in filling the void due to the lack of protein level information since they provide information on the substrates and products of proteins, specifically enzymes.

As of now, over 40,000 metabolites have been identified in human cells according to the Human Metabolome Database (HMDB) (Wishart et al. 2007, 2009, 2013). These metabolites can be intermediates or products of cellular metabolism, which include the basic metabolites such as amino acids, nucleotides, alcohols, organic acids and vitamins, and complex metabolites such as cholesterol and steroid hormones. By analyzing the quantitative data of metabolites associated with a specific metabolic pathway, it is possible to make a generally accurate estimate of the

activity level of the pathway. For example, glucose-6-phosphate, fructose-6-phosphate, glyceraldehyde 3-phosphate, phosphoenol-pyruvate, pyruvate and lactate are the main metabolites of the glycolytic fermentation pathway (see Fig. 1.4), and their relative abundances provide accurate information about the activity level of the pathway. By carrying out metabolite flux analyses (Varma and Palsson 1994) based on the pathway information and the measured quantities of these metabolites, one can infer if some of these intermediates or end products may be directed towards other metabolic pathways, in addition to being part of the glycolytic pathway.

Metabolic flux analysis generally applies to any well-established biological pathway, such as those in central metabolism. That is, with all the relevant reactions and the encoding genes known, metabolic data can be used in conjunction with the matching transcriptomic data, to infer the flux of a specific molecular species such as carbon or nitrogen. In essence, this provides flux information of different elements across an entire network, which preserves balances between the total input and the output elements for each reaction, hence providing a systems-level representation of the flux distribution across all the branch points in the network. Identification of unbalanced reactions, i.e., the total number of carbons into a reaction is different from that out of the reaction, can help to detect previously unknown branches involved in the relevant reactions. This type of analysis can be used to identify possible relationships between two known metabolic pathways, such as detecting possible metabolic relationships between the glutaminolysis pathway (McKeehan 1982), which tends to be up-regulated in cancer cells, and other metabolic pathways, or detection of relationships between cholesterol metabolism and phospholipid metabolism in metastatic cancer (see Chap. 11). For example, an analysis like this has led us to detect that some metabolites of the glycolysis pathway become substrates of another metabolic pathway, the *hyaluronic acid synthesis* pathway (see Chap. 6). When reaction rate constants are available or can be estimated for all the relevant enzymes, one can identify the rate-limiting steps in a pathway, thus enabling one to undertake detailed mechanistic studies of a biological process.

Both high-resolution mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy have been used to identify metabolites present in cells and in tissue samples, each having their own advantages and limitations. MS can provide quantitative measures for up to 1,000 different metabolite species, but it suffers from relatively low repeatability (Boshier et al. 2010). In comparison, NMR can provide highly accurate measurements of metabolites but is limited in the number of metabolite species in each experiment. With either type of instrument, one can obtain quantitative measures of numerous metabolite species.

When coupled with transcriptomic data and functional annotations of genes, metabolomic data can be used to infer the detailed metabolic pathway that may produce specific metabolites. Specifically, for each experimentally identified metabolite in a sample, one can search for enzymes among the expressed enzyme-encoding genes that may be responsible for its synthesis through comparisons against the Enzyme Classification (Bairoch 2000) or KEGG database. Both of these databases contain information about enzymes and substrates that can lead to the production of

a specific metabolite. If there is more than one candidate, a selection can be inferred by finding the one that is most consistent with the available transcriptomic and functional annotation data, i.e., the enzyme-encoding gene is expressed and the substrate is among the identified metabolites. By repeating this process, one can create a pathway consisting of the identified enzymes along with the identified metabolites. Although one cannot expect to derive all the relevant enzymes along the pathway, it is generally possible to develop a crude model based on our own experience. In addition, it is possible to expand a pathway model through careful applications of the transcriptomic data and the metabolomic data collected to identify previously unknown or poorly studied branches of well-studied pathways. For example, by carefully analyzing the metabolites associated with glycolysis, one can possibly identify those that serve as intermediates between glycolytic metabolites and metabolites involved in the synthesis of hyaluronic acids as detailed in Chap. 6.

There are a number of databases for human metabolomic data in the public domain, including the HMDB, BiGG (metabolic reconstruction of human metabolism) (Schellenberger et al. 2010) and the Tumor Metabolism database (The-Tumor-Metabolome 2011). Another useful database is Brenda (Scheer et al. 2011), which provides the reaction parameters of various enzymes. All these databases provide useful information needed for reconstruction of specific metabolic processes in normal and cancer cells.

2.5 Patient Data

Knowledge of patient data is essential for the correct interpretation of their respective *omic* data. People of different gender, age and race, and with different histories of smoking, alcohol consumption and health problems, could have different baseline gene-expression levels. It was noted, from our previous studies, that some genes are sensitive to one aspect of a person's attributes, such as age or gender, while other genes may be more sensitive to other attributes. And some genes are attribute-independent. For example, based on our analysis on gene-expression data of 80 gastric cancer tissues and their matching tissues from 80 patients (see Appendix of Chap. 3 for details of the dataset), it was found that the expression levels of some genes are age-dependent, gender-dependent and smoking history-dependent, while other genes are, in large part, independent of any of these features (Cui et al. 2011). When working with these datasets, it was noted that the baseline expression levels of 143 genes were highly age-dependent, including *MUC1* (mucin 1), *UBFD1* (ubiquitin family domain 1) and *MDK* (neurite growth-promoting factor 2). In addition, 59 genes were gender-dependent; these included *WNT2* (wingless-type MMTV integration site family, member 2), *ARSE* (arylsulfatase E) and *KCNN2* (potassium intermediate/small conductance calcium-activated channel, subfamily N, member 2) (see (Cui et al. 2011) for details). Similar analyses can be carried out on dependence using various lifestyle habits such as smoking and medications.

Knowing such information, one then needs to make age or gender corrections on the observed gene-expression data before interpreting the data for functional inference. The detailed correction scheme depends on the actual relationship between a specific attribute and the gene-expression levels. Various normalization techniques and software tools are publicly available for this purpose.

2.6 A Case Study of Integrative *Omic* Data Analyses

We present an example here to show how integrative analyses of multiple *omic* and computational data types can lead to new insights about cancer mechanisms. The main question being addressed here is: *What makes metastatic cancers grow substantially faster than their primary cancer counterparts*? While a detailed model for this problem is given in Chap. 11, the current focus is on how this problem can be approached through transcriptomic data analyses coupled with limited metabolomic data analyses.

To address this question, all the transcriptomic data of metastatic cancers, along with their corresponding primary cancers, were collected on the Internet. Sixteen large sets of genome-scale transcriptomic data covering 11 types of metastatic and corresponding primary cancers were extracted from the GEO database, including breast to bone, breast to brain, breast to liver, breast to lung, colon to liver, colon to lung, kidney-to-lung, pancreas to liver and lung, and prostate to bone and liver. The detailed information of these datasets is given in Chap. 11.

The first question addressed is: *Which genes are consistently up-regulated in metastatic cancers in comparison with their corresponding primary cancers across all these datasets*? Simple statistical analyses led to the identification of about 100 such genes.

The second question asked is: *What do these genes do in terms of cellular function(s)*? Pathway enrichment analyses of these genes using DAVID against KEGG, REACTOME and BIOCARTA revealed that the most significantly enriched pathway was "cholesterol uptake and metabolism". Two questions were then asked: (a) *What does cholesterol do in metastatic cancer cells*? And (b) *Why do metastatic cancer cells need more cholesterol*, as suggested by the observation that at least one cholesterol-containing lipoprotein transporter gene, *SRB1* (scavenger receptor B), *LDLR* (low density lipoprotein receptor) or *VLDLR* (very low density lipoprotein receptor) was substantially up-regulated compared to the corresponding primary cancers except for some brain metastases. These metastases synthesize cholesterol *de novo* as cholesterol-containing lipoproteins probably could not enter brain tissue due to the blood-brain barrier (Bjorkhem and Meaney 2004).

Here, only the first question is considered. It was noted that multiple *CYP* (cytochrome P450) genes are up-regulated in each metastatic cancer type: these genes encode enzymes for oxidizing cholesterols to various oxysterols or bile acids. Some of these oxysterols are further metabolized to steroid hormones such as estrogens,

androgens or steroidogenic derivatives by various enzymes whose genes show substantially increased expression levels in comparison with their corresponding primary cancers. A number of these steroid products can bind with and activate different nuclear receptors, such as *FXR* (farnesoid X receptor) and *ER* (estrogen receptor) (see Chap. 11). Various growth-factor receptors such as *FGFR* (fibroblast growth factor receptor) and *EGFR* (epidermal growth factor receptor) are up-regulated in different metastatic cancers, some of which can be directly activated by oxysterols and/or steroid hormones, whose abundances tend to be substantially elevated in metastatic cancers. For the other growth factor receptors, strong correlations between their gene-expressions and the expression patterns of the various nuclear receptors are observed across different metastases, thus suggesting the possibility of a functional relationship between the activation of the two sets of receptors. Based on more detailed analyses and validation, a mechanistic model for how metastatic cancers utilize oxidized cholesterols to accelerate their growth is presented in Chap. 11. Similar integrative analyses of multiple types of data can be carried out to derive the mechanistic models for a large variety of poorly understood cancer-related processes if one can ask the right questions that could be answered through analyses and mining of the relevant *omic* data.

2.7 Concluding Remarks

A substantial amount of information concerning the activities of individual biochemical pathways, their dynamics and the complex relationships among them, and with respect to various micro-environmental factors, is hidden in the very large pool of publicly available cancer *omic* data, including transcriptomic, genomic, metabolomic and epigenomic data. Powerful statistical analysis techniques can aid immensely in uncovering such information if one poses the right questions. Such focused questions create a framework for hypothesis-guided data analysis and mining to check for the validity of the formulated hypothesis, as well as for guiding the formulation of further questions, which may ultimately lead to the elucidation of specific pathways or even possibly causal relationships among the activities of different pathways. More powerful analysis tools for different *omic* data types are clearly needed in order to address more complex and deeper questions about the available data such as de-convolution of gene-expression data collected on tissue samples consisting of multiple cell types and inference of causal relationships. Integrative analyses of multiple types of *omic* and computational data will prove to be the key to effective data mining and information discovery. A large number of examples are presented throughout the following chapters regarding how best to address various cancer biology inquiries, including fundamental questions, through mining the available *omic* data.

References

Bairoch A (2000) The ENZYME database in 2000. Nucleic acids research 28: 304–305
Barkow S, Bleuler S, Prelic A et al. (2006) BicAT: a biclustering analysis toolbox. Bioinformatics 22: 1282–1283
Bentley J, Diggle CP, Harnden P et al. (2004) DNA double strand break repair in human bladder cancer is error prone and involves microhomology-associated end-joining. Nucleic acids research 32: 5249–5259
Bindra RS, Crosby ME, Glazer PM (2007) Regulation of DNA repair in hypoxic cancer cells. Cancer Metastasis Rev 26: 249–260
Bjorkhem I, Meaney S (2004) Brain cholesterol: long secret life behind a barrier. Arterioscler Thromb Vasc Biol 24: 806–815
Boshier PR, Marczin N, Hanna GB (2010) Repeatability of the measurement of exhaled volatile metabolites using selected ion flow tube mass spectrometry. J Am Soc Mass Spectrom 21: 1070–1074
Cancer-Genome-Atlas-Research (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455: 1061–1068
Cancer-Genome-Atlas-Research (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–615
Cancer-Genome-Atlas-Research (2012a) Comprehensive genomic characterization of squamous cell lung cancers. Nature 489: 519–525
Cancer-Genome-Atlas-Research (2012b) Comprehensive molecular characterization of human colon and rectal cancer. Nature 487: 330–337
Cancer-Genome-Atlas-Research (2012c) Comprehensive molecular portraits of human breast tumours. Nature 490: 61–70
Cancer-Genome-Atlas-Research (2013a) Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499: 43–49
Cancer-Genome-Atlas-Research (2013b) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. The New England journal of medicine 368: 2059–2074
Cancer-Genome-Atlas-Research, Weinstein JN, Collisson EA et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nature genetics 45: 1113–1120
Cavalli-Sforza LL (2005) The Human Genome Diversity Project: past, present and future. Nat Rev Genet 6: 333–340
Chen JS, Hung WS, Chan HH et al. (2013) In silico identification of oncogenic potential of fyn-related kinase in hepatocellular carcinoma. Bioinformatics 29: 420–427
Cheng Z, Ventura M, She X et al. (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437: 88–93
Church GM (2005) The personal genome project. Mol Syst Biol 1: 2005 0030
Colotta F, Allavena P, Sica A et al. (2009) Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability. Carcinogenesis 30: 1073–1081
Cotton RG, Auerbach AD, Axton M et al. (2008) GENETICS. The Human Variome Project. Science 322: 861–862
Croft D, O'Kelly G, Wu G et al. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic acids research 39: D691–697
Cui J, Chen Y, Chou WC et al. (2011) An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer. Nucleic acids research 39: 1197–1207
D'Souza I, Poorkaj P, Hong M et al. (1999) Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements. Proceedings of the National Academy of Sciences of the United States of America 96: 5598–5603
Downing JR, Wilson RK, Zhang J et al. (2012) The Pediatric Cancer Genome Project. Nature genetics 44: 619–622

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research 30: 207–210

Emmert-Buck MR, Bonner RF, Smith PD et al. (1996) Laser capture microdissection. Science 274: 998–1001

Essaghir A, Toffalini F, Knoops L et al. (2010) Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. Nucleic acids research 38: e120

Forbes SA, Bindal N, Bamford S et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research 39: D945–D950

Grabherr MG, Haas BJ, Yassour M et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644–652

Green RE, Krause J, Briggs AW et al. (2010) A draft sequence of the Neanderthal genome. Science 328: 710–722

Greenman C, Stephens P, Smith R et al. (2007) Patterns of somatic mutation in human cancer genomes. Nature 446: 153–158

Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144: 646–674

Huang D, Sherman BT, Tan Q et al. (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 8: R183

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 4: 44–57

Huebert DJ, Kamal M, O'Donovan A et al. (2006) Genome-wide analysis of histone modifications by ChIP-on-chip. Methods 40: 365–369

International-HapMap (2003) The International HapMap Project. Nature 426: 789–796

Iyer VR, Horak CE, Scafe CS et al. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409: 533–538

Kamakaka RT, Biggins S (2005) Histone variants: deviants? Genes & development 19: 295–310

Kandoth C, Schultz N, Cherniack AD et al. (2013) Integrated genomic characterization of endometrial carcinoma. Nature 497: 67–73

Kanehisa M, Goto S, Furumichi M et al. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic acids research 38: D355–360

Kanehisa M, Goto S, Sato Y et al. (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research 40: D109–114

Kanehisa M, Goto S, Sato Y et al. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic acids research 42: D199–205

Lander ES, Linton LM, Birren B et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921

Li G, Ma Q, Tang H et al. (2009) QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. Nucleic acids research 37: e101

Liu R, Hannenhalli S, Bucan M (2009) Motifs and cis-regulatory modules mediating the expression of genes co-expressed in presynaptic neurons. Genome Biol 10: R72

Lu P, Weaver VM, Werb Z (2012) The extracellular matrix: a dynamic niche in cancer progression. The Journal of cell biology 196: 395–406

McKeehan W (1982) Glycolysis, glutaminolysis and cell proliferation. Cell Biology International Reports 6: 635–650

Nazarov PV, Muller A, Khutko V et al. Co-Expression Analysis of Large Microarray Data Sets using Coexpress Software Tool. In: Seventh International Workshop on Computational Systems Biology, WCSB 2010, 2010.

Nishimura D (2001) BioCarta. Biotech Software & Internet Report 2:

Nowell P, Hungerford D (1960) A minute chromosome in human chronic granulocytic leukemia. Science 132:

Pearl J (2009) Causal inference in statistics: An overview. Statistics Surveys 3: 96–146

Qian J, Lin J, Luscombe NM et al. (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics 19: 1917–1926

Ren B, Robert F, Wyrick JJ et al. (2000) Genome-wide location and function of DNA binding proteins. Science 290: 2306–2309

Scanlon S, Glazer P (2013) Genetic Instability Induced by Hypoxic Stress. In: Mittelman D (ed) Stress-Induced Mutagenesis. Springer New York, pp 151–181

Scheer M, Grote A, Chang A et al. (2011) BRENDA, the enzyme information system in 2011. Nucleic acids research 39: D670–676

Schellenberger J, Park JO, Conrad TM et al. (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11: 213

Segal E, Yelensky R, Kaushal A et al. (2004) GeneXPress: A Visualization and Statistical Analysis Tool for Gene Expression and Sequence Data. Paper presented at the 11th International Conference on Intelligent Systems for Molecular Biology (ISMB),

Service RF (2006) Gene sequencing. The race for the $1000 genome. Science 311: 1544–1546

Steel M, Penny D (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. Molecular biology and evolution 17: 839–850

Strahl BD, Allis CD (2000) The language of covalent histone modifications. Nature 403: 41–45

The-National-Cancer-Act (1971) The National Cancer Act of 1971.

The-Tumor-Metabolome (2011) The tumor metabolome.

Trapnell C, Williams BA, Pertea G et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511–515

Truong LN, Li Y, Shi LZ et al. (2013) Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. Proceedings of the National Academy of Sciences of the United States of America 110: 7720–7725

Valen LV (1973) A new evolutionary law. Evolutionary Theory 1: 1–30

Varma A, Palsson BO (1994) Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. Nature Biotechnology 12: 994–998

Venter JC, Adams MD, Myers EW et al. (2001) The sequence of the human genome. Science 291: 1304–1351

Vignali M, Hassan AH, Neely KE et al. (2000) ATP-dependent chromatin-remodeling complexes. Molecular and cellular biology 20: 1899–1910

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63

Wishart DS, Jewison T, Guo AC et al. (2013) HMDB 3.0–The Human Metabolome Database in 2013. Nucleic acids research 41: D801–807

Wishart DS, Knox C, Guo AC et al. (2009) HMDB: a knowledgebase for the human metabolome. Nucleic acids research 37: D603-610

Wishart DS, Tzur D, Knox C et al. (2007) HMDB: the Human Metabolome Database. Nucleic acids research 35: D521–526

Xu K, Mao X, Mehta M et al. (2012) A comparative study of gene-expression data of basal cell carcinoma and melanoma reveals new insights about the two cancers. PLoS One 7: e30750

Yang AS, Estecio MR, Doshi K et al. (2004) A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements. Nucleic acids research 32: e38

Zhao M, Sun JC, Zhao ZM (2013) TSGene: a web resource for tumor suppressor genes. Nucleic acids research 41: D970–D976