Chapter 3
Cancer Classification and Molecular Signature Identification

Cancer is a family of diseases that share a common set of characteristics such as reprogrammed energy metabolism, uncontrolled cell growth, tumor angiogenesis and avoidance of immune destruction, referred to as cancer hallmarks, as introduced in Chap. 1. Based on their original cell types, cancers are classified into five classes: (1) *carcinoma*, which begins in epithelial cells and represents the majority of the human cancer cases; (2) *sarcoma*, derived from mesenchymal cells, e.g., connective tissue cells such as fibroblasts; (3) *lymphoma, leukemia* and *myeloma*, originating in hematopoietic or blood-forming cells; (4) *germ cell tumors*, developing, as the name implies, from germ cells; and (5) *neuroblastoma*, *glioma, glioblastoma* and others derived from cells of the central and peripheral nervous system and denoted as *neuroectodermal* tumors because of their beginning in the early embryo. Each class may consist of cancers of different types. For example, carcinoma comprises adenocarcinoma, basal-cell carcinoma, small-cell carcinoma and squamous cell carcinoma, independent of their underlying tissue types. Cancers of the same type and developing in the same tissue may have distinct properties in terms of their growth patterns, malignance levels, survival rates and possibly even different underlying mechanisms. They may respond differently to the same drug treatment and hence have different mortality rates. As of now, over 200 types of human cancers have been identified and characterized (Stewart and Kleihues 2003), the majority of which are determined based on the location, the originating cell type and cell morphology. It is now becoming evident that this type of classification, in large part subjective, is not adequate for developing personalized treatment plans, which are becoming increasingly desirable and clearly represents the future of cancer medicine.

With the rapid accumulation of high-throughput *omic* data for cancer, particularly transcriptomic and genomic data, it is now feasible to classify cancers based on their molecular level information. For example, this can be based on distinct expression patterns of certain genes or pathways shared only by samples of the same cancer

Y. Xu et al., *Cancer Bioinformatics*, DOI 10.1007/978-1-4939-1381-7_3

type, or combinations of mutations that tend to co-occur (or be selected, to be more accurate) in certain cancer types. Such type-defining expression or mutation patterns of genes are referred to as the *signature* of a cancer type. This idea should be applicable to every kind of cancer as has been done for a few cancer types, such as Oncotype DX for one form of breast cancer (Albain et al. 2010), as long as transcriptomic or genomic mutation data are available for the cancer category. Similarly, it should also be possible to derive molecular signatures for cancer grades and cancer stages, with the former referring to the level of malignancy of a tumor and the latter representing the location of the cancer in its development towards the terminal stage, i.e., metastasis. Compared to the traditional definitions of cancer types, molecular signatures, as outlined here, can potentially provide more accurate characterization of a cancer and even reveal its underlying mechanisms, hence possibly having significant implications to cancer treatment and prognosis prediction. Here we use gene-expression data as an example to illustrate how cancer typing, staging and grading can be done using *omic* data, which could potentially lead to substantially more accurate characterization of cancers of different types, grades and stages. Similar ideas should be applicable to mutation-based cancer classification.

3.1 Cancer Types, Grades and Stages

The earliest description of cancer can be traced back to 2500 BC by Egyptian physician Imhotep (Mukherjee 2010). Evidence exists suggesting that Egyptian physicians at the time could distinguish between benign and malignant tumors. The study of cancer as a scientific discipline came in the nineteenth century when microscopes became widely available to physicians and surgeons. Microscopic pathology, pioneered by German doctor Rudolf Virchow, laid the foundation for the development of cancer surgery as practiced now. Since then, cancer tissues removed from patients are microscopically examined and classified based on their morphological characteristics. Scientific oncology was born out of the debate concerning a few competing hypotheses regarding the possible causes of cancer in the late 1800s through the early 1900s. It developed based on findings that linked microscopic observations made on cancer tissues to clinical data during the course of the disease development. The popular hypotheses included: (1) one proposed by Stahl and Hoffman, which suggested that cancer was caused by coagulated lymph; (2) a proposal by Johannes Muller who suggested that cancer cells arose from budding elements between normal tissues; and (3) the theory developed by Rudolph Virchow, which considered cancer as a disease of cells. The next major advance in attempts to elucidate the possible causes of a cancer came in the 1920s when the German biochemist Otto Warburg observed that cancer cells rely heavily on glycolytic fermentation rather than the more efficient oxidative phosphorylation for ATP generation, even when oxygen is available. This metabolic alteration is referred to as the *Warburg effect* (Warburg 1956) and remains under active investigation as discussed in depth in Chap. 5. Based on the accelerated glycolysis, some 10 to 20-fold over that

of normal cells, Warburg attributed cancer to a malfunctioning mitochondria-induced metabolic disease. The discovery of oncogenes in 1970s by Bishop and Varmus, along with the discovery of tumor-suppressor genes by A. G. Knudson also in 1970s, represented the next key advancement, which started the era of classifying cancer as a genetic disease.

Early classification of cancers was based on a cancer's location, such as lung cancer, skin cancer or blood cancer (e.g., leukemia). Over time, oncologists began to realize that different types of cancers can develop from the same organ. The earliest classification of cancers from the same organ, in this case bone marrow which houses the hematopoietic stem cells, can be traced back to the early 1900s when it was found that there were at least four types of leukemia, namely ALL (acute lymphoblastic leukemia), AML (acute myelogenous leukemia), CLL (chronic lymphoblastic leukemia) and CML. This realization occurred about 50 years after the diagnosis of the first documented leukemia case (Beutler 2001). For other cancers, recognition of multiple cancer types originating from the same organ came rather late. For example, small-cell lung cancer was not considered as a separate type of lung cancer from the more prevalent and less aggressive non-small cell lung cancer until the 1960s. Gastric cancers were found to have at least two subtypes, intestinal and diffuse, in 1965 (Lauren 1965). It is worth noting that correct diagnosis of a cancer type has significant implications to designing the most effective treatment protocols and prognosis. For example, statistics show that the current 5-year survival rates for adult ALL, AML, CLL and CML patients are 50 %, 40 %, 75 % and 90 %, respectively, and the treatment plan for each of them is quite different. ALL is typically treated using chemotherapy followed by anti-metabolite drugs; AML is generally treated using chemotherapy; CLL, while incurable, is often being controlled with chemotherapy using a combination of fludarabine and alkylating agents; and CML is, in most cases, successfully treated using the so called "miracle" drug Gleevec, or else newer and improved drugs.

The multistage nature of a cancer was first discovered by Japanese researchers Yamagiwa and Ichikawa in the beginning of the twentieth century (Yamagiwa and Ichikawa 1918). Basically for most cancer types, the histological stage refers to the extent the cancer has spread, which is typically numbered from stage I through stage IV, with IV representing the most advanced stage. The stage of a cancer is an important predictor for survival, with the treatment plan often determined based on staging. Currently the stage of a cancer is generally determined by pathological analysis from a biopsied specimen of the cancer tissue, including lymph nodes, as well as analysis by imaging techniques with the results interpreted by radiologists; only limited molecular level information such as the expression levels of a few marker genes as determined by immune-detection.

In addition to type and stage, cancer grade is another important parameter that has been used by pathologists to represent the level of malignancy of a given cancer, determined based on surgical specimens. This parameter is largely independent of the type and the stage of a cancer. A popular grading system uses four grades: (1) G1 (highly differentiated), (2) G2 (moderately differentiated), (3) G3 (poorly differentiated) and (4) G4 (undifferentiated), with G4 representing the most malignant.

The level of differentiation refers to the maturity of a cell in developmental biology. In the current context, the more differentiated cancer cells resemble more of the normal mature cells, and they tend to grow and spread at slower rates than undifferentiated or poorly differentiated cancer cells. The grade of a cancer provides another key indicator for prognosis. While the term seems to be defined in terms of cellular differentiation, the actual determination of the cancer grade is often made based on a combination of the cellular appearance (degree of abnormality), the rate of growth and the degree of invasiveness.

The current availability of significant quantities of molecular level *omic* data on cancer, such as transcriptomic, genomic, epigenomic and metabolomic data, provides unprecedented opportunities for developing molecular-level signatures for each known cancer type, grade and stage, and, if needed, possibly reclassifying some of the previously determined cancer types, stages and/or grades. This has the potential to lead to more accurate classifications of a cancer for the purpose of improved treatment design and prognosis evaluation.

3.2 Computational Cancer Typing, Staging and Grading Through Data Classification

The main question addressed here is: For a given set of cancer samples, each marked with a specific type, stage or grade determined by pathologists*, is it possible to identify common characteristics,* e.g.*, in terms of gene expression patterns among samples having the same class label*? If the answer is yes, such a capability could potentially be used to accurately define the type or subtype, stage or substage, grade or subgrade of a cancer. In the following sections, we demonstrate how this could be done to possibly provide a new way of classifying cancer based on molecular level data.

3.2.1 Cancer Typing

A basis for gene-expression data-based cancer typing is that cancers of various types have their distinct phenotypic characteristics such as differences in cellular shape, growth rates and responses to the same treatment regiments, and possibly distinct underlying mechanisms, while samples of the same type tend to share common characteristics. These phenotypic and mechanistic commonalities among cancer cases of the same type as well as differences across multiple cancer types are realized through molecular level activities and hence should be in general reflected by the expression patterns of some genes. A key in accomplishing cancer typing based on gene-expression data is to identify those genes whose expression patterns are shared by samples of the same type but not shared by samples of the other cancer types. This problem can be modeled computationally in various ways,

depending on the specific purpose(s) of the cancer typing. For example, if the goal is to identify the defining characteristics of a cancer type, one may decide to identify a maximal gene set, whose expression patterns are similar across all the (available) cancer samples of the same type and different from those of other types. If, instead, the goal is to identify distinguishing characteristics between two (or more) types of cancers, one may want to find a minimal set of genes whose expression patterns can delineate among samples between the two (or more) cancer types, which may not necessarily contain any information about the distinct mechanisms of the different cancer types.

We now present one example to model the cancer typing problem and to illustrate how such a problem can be solved computationally. Consider two subtypes of gastric cancer, the intestinal (C_1) and diffuse (C_2) subtypes, each having genome-scale gene-expression data collected using the same platform on paired cancer and matching control tissue samples from the same patients. For each patient one can obtain the fold-change information for any gene between its expression in a cancer and its matching control, which is typically calculated as the logarithm of the ratio between the two expression levels, referred to as the *log-ratio* throughout this book. The present goal is to find a minimal subset of genes out of the total of ~20,000 human genes, whose expression patterns can unequivocally distinguish between the two subtypes, C_1 and C_2. Specifically, the aim is to identify a set G of genes and a discriminant function F() so that $F(G(x))>0$ for $x \in C_1$ and $F(G(x))<0$ for $x \in C_2$ for as many $x \in C_1 \cup C_2$ as possible, where G(x) represents the list of fold-changes in expression levels of genes in G between cancer tissue x and its matching control. There are many classes of discriminant functions that can be used for solving this classification problem. Here a specific class of functions is used, the linear *support vector machine* (SVM) (Cortes and Vapnik 1995). The goal now becomes that of locating a minimal set G of genes and an SVM that achieve the best classification with the misclassification rate lower than a pre-defined threshold δ.

One method of solving this problem is by going through all combinations of K genes among all the human genes, searching from $K=1$ and up until an SVM-based classifier and a K-gene set G are found, which achieve the desired classification accuracy defined by δ. In practice, the search will not include all the human genes since the majority will not be expressed for any specific tissue type. For this problem, one only needs to consider genes that are differentially expressed between cancer samples and the matching controls. To get a sense of the amount of computing time that may be needed to exhaustively search through all K-gene combinations, consider the following typical scenario: the two gene-expression datasets with C_1 having 100 pairs of samples and C_2 consisting of 150 pairs of samples; and 500 genes showing differential expressions (see Chap. 2) between the two sets of samples. In this case, one would need to examine $\binom{500}{K}$ combinations to find a K-gene combination that achieves the optimal classification between the two datasets. For each K-gene combination, a linear SVM is trained to optimally classify the two datasets as discussed above; if a trained SVM achieves a classification accuracy better than δ, retain the SVM as a candidate classifier; then repeat this process until all

K-gene combinations are exhausted. The final classifier is the one with the lowest misclassification rate among all those retained. Our experience has been that *K* should be no larger than 8; otherwise the number $\binom{500}{K}$ may be too large for a desktop workstation to handle. The following gives a detailed procedure of the search process:

Cancer classification algorithm

FOR $K = 1$ **TO** N **DO**

FOR each *K*-gene combination from the pool of differentially expressed genes **DO**

a. **DO** the following **FOR** 1,000 times

1. Randomly split C_1 and C_2 into C_1-training and C_1-testing, and C_2-training and C_2-testing, respectively, with C_x-training and C_x-testing having the same size, $x \in \{1, 2\}$;
2. Train a linear SVM based on the current *K*-gene combination on C_1-training and C_2-training, which achieves optimal classification between C_1-testing and C_2-testing;
3. **IF** the misclassification rate of the trained SVM is $< \delta$, **THEN** keep the SVM;

b. **IF** at least one SVM for the *K*-gene combination has misclassification rates $< \delta$, **THEN** keep the *K*-gene combination with the lowest misclassification rate a candidate for the final classifier.

IF at least one final classifier candidate is found, **THEN OUTPUT** the one with the lowest misclassification rate, **ELSE OUTPUT** no classifier is found with at most *N* genes and misclassification rate $< \delta$.

where *N* is the upper bound (set by the user) for searching a satisfying *K*-gene discriminator, and 1,000 is the number of times used to find an optimal *K*-gene classifier over different partitions of the given datasets C_1 and C_2.

This simple procedure has been used to find an optimal SVM-based classifier between the two subtypes of gastric cancer based on gene-expression data collected on 80 pairs of gastric cancer and matching controls (Cui et al. 2011a). Figure 3.1 shows classification accuracies by the best *K*-gene classifiers for $K \leq 8$.

If one needs to search for a *K*-gene classifier with larger *K*'s (>8) for some application, a different search strategy may be needed to make it computationally feasible. One such strategy is called *recursive feature elimination*, a procedure often used in conjunction with an SVM application; together they are referred to as *RFE-SVM*. While the detailed information of an RFE-SVM procedure can be found in (Guyon et al. 2002; Inza et al. 2004), the basic idea is to start with a list of all genes, each having some discerning power in distinguishing between the two classes of samples, and to train a classifier, followed with the RFE procedure to repeatedly

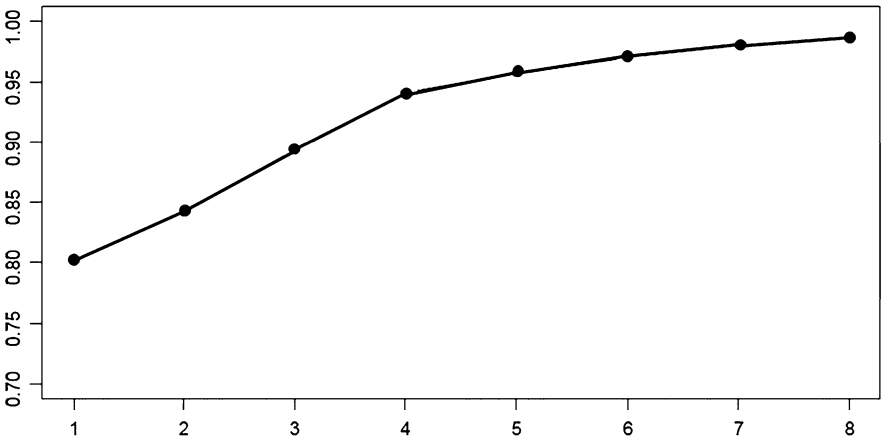

Fig. 3.1 SVM-based classification accuracy using the best K-gene combination, for $K = 1, 2, \ldots, 8$, on 80 pairs of gastric cancer and control tissues

remove genes from the initial gene list as long as the classification accuracy is not affected until only K genes are left.

If desired, this idea for solving a 2-class classification problem can be generalized to M-class problems, for $M > 2$, so multi-type cancers originating from the same tissue, such as the different types of leukemia, can be classified based on identification and application of K-gene combinations as done above. One specific way to accomplish this is given as follow: a M-class classifier can be constructed by separately calculating M binary classifiers, each separating class i from the remaining classes, $i = 1, \ldots, M$. Then, an input sample is classified to class J if the sample has the highest classification significance by the J^{th} classifier. Such a method is regarded as one-*versus*-all multi-class SVM (Cui et al. 2011a). A detailed review on such classifiers can be found in (Duan and Keerthi 2005). Using this type of classification method, one can build classifiers for all the cancer types as long as they have gene-expression data available, along with labeled type information for each sample.

Numerous K-gene combinations, also referred to as *K-gene panels*, have been identified and used as signatures for various cancer types. For example, a panel of 104 genes has been identified for distinguishing cancer tissues (of multiple types) from healthy tissues (Starmans et al. 2008), aimed to detect if a tissue is cancerous or not. Other signature panels include: (1) a 70-gene panel for predicting the potential for developing breast cancer, built by MammaPrint (Slodkowska and Ross 2009); (2) a 21-gene panel, termed *Oncotype DX*, for a similar purpose; (3) a 71-gene panel for identification of cancers that are sensitive to *TRAIL*-induced apoptosis (Chen et al. 2012); (4) a 31-gene panel used to predict the metastasis potential of a breast cancer, developed by CompanDX (Cho et al. 2012); and (5) a 16-gene panel for testing for non-small-cell lung cancer against other lung cancer

types (Shedden et al. 2008). Having a test kit for a specific cancer type, e.g., metastasis-prone or not, can enable surgeons to make a rapid and informed decision regarding the appropriate surgical procedure to adopt. Other test kits can assist oncologists in making an informed decision regarding the most appropriate treatment plan for a particular cancer case. For example, *TRAIL* (*TNF*-related apoptosis inducing ligand) is an anticancer-mediating protein that can induce apoptosis in cancer cells but not in normal cells. This makes *TRAIL* highly desirable; however, not all cancers are sensitive to *TRAIL*. Hence, having a test using such a kit can quickly determine if a cancer patient should be treated with *TRAIL* or not.

In order to ensure the general applicability of any identified signature genes, it is essential to carry out proper normalization of the to-be-used transcriptomic data that may be collected by different research labs, specifically to correct any systematic errors in the data caused by different sample-preparation and data-collection protocols. Batch-based normalization such as the model presented in (Johnson et al. 2007) may prove to be effective in removing so created systematic errors due to using different data-collection protocols.

Although a number of computational methods have been developed for defining cancer types using gene-expression data (Ramaswamy et al. 2001; Tibshirani et al. 2002; Weigelt et al. 2010; Reis-Filho and Pusztai 2011), none of them have achieved 100 % consistency with the typing results determined by cancer pathologists. There may be two key reasons for the less-than-perfect agreement. One is that some of the cancer typing decisions by pathologists may not necessarily be correct for various reasons: (a) a cancer identification protocol may use only limited molecular level and somewhat subjective visual information; and (b) there is always the possibility of human errors in executing a type-calling procedure, particularly when visual appearances may be borderline between different options. Another possibility could be due to limitations of the current classification techniques. For example, the above classification methods may be too simple to capture the complex relationships among the expression data of multiple genes, which are unique to a specific cancer type. Moreover, it may be due to something more fundamental, such as the gene expression data not necessarily having all the information needed to classify cancer types correctly, e.g., some of the needed information may be at the protein or the post-translational level. It is expected that answers to this question may emerge as more cancer *omic* data become available and/or when more advanced analysis techniques will be developed.

3.2.2 Cancer Staging

Cancer stages have been defined mainly in terms of the tumor size, cell morphology and the state of metastasis. Currently its determination involves some level of subjectivity by pathologists. Like cancer types, cancer stages can also be defined in terms of expression patterns of some subset of the human genes. A number of studies have been published on applications of computational techniques to predict the

stage of a cancer based on gene-expression data (Eddy et al. 2010; Goodison et al. 2010; Liong et al. 2012). For example, a 7-gene panel (*ANPEP, ABL1, PSCA, EFNA1, HSPB1, INMT, TRIP13*) was used to measure the progression of prostate cancer and achieved high-80 % consistencies with pathologically-determined stages (Liong et al. 2012). Another example is a 4-gene panel (*IL1B*, *S100A8*, *S100A9*, *EGFR*) for assessing the progression of muscle invasive bladder cancer (Kim et al. 2011). Similar gene panels have been developed for a few other cancers, such as breast cancer (Rodenhiser et al. 2011; Arranz et al. 2012), colon cancer (Erten et al. 2012) and oral cancer (Mroz and Rocco 2012).

Potentially, one can develop such gene-panels for any cancer as long as transcriptomic data for cancer and control tissues, along with their stage information, are available. Here we use gastric cancer again as an example to illustrate how gene-expression data can be used to predict the developmental stage of a cancer.

The same set of gene-expression data collected on 80 pairs of gastric cancer and matching noncancerous gastric tissues used in Sect. 3.2.1 is again analyzed. Of the 80 cancer tissues, 4 were in stage I, 7 in stage II, 54 in stage III and 15 in stage IV. The detailed gene-expression data of these samples can be found in the Appendix. Note that these tissue samples are not evenly distributed across the four stages, but this may be a good representation of the actual stage distribution for gastric cancer patients presenting for resection, at least in China where the 80 samples were collected. The present goal is to identify a set of differentially expressed genes between cancer and the matching controls, where the expression patterns adequately reflect the stages of all the gastric cancer samples. On this data set of 80 pairs of samples, 715 genes were found consistently to be differentially expressed between the cancer and the matching controls (Cui et al. 2011a).

A simplified version of the staging problem is considered first, by merging stages I and II samples into one "early stage" group and stages III and IV samples into the "advanced stage" group, making this a 2-stage classification problem. From an analysis of all the differentially-expressed genes, four genes, *CHRM3* (cholinergic receptor), *PCDH7* (protocadherin), *SATB2* (special AT-rich sequence-binding protein) and *PPA1* (pyrophosphatase), were identified, each giving a consistency level with the two combined stages better than 80 % by using a simple fold-change cut-off. When using K-gene combinations for $K>1$, the classification consistency (with pathologist-determined stages) continues to increase as K increases until it reaches 95 %, and then the improvement becomes asymptotic.

Using the generalized classification scheme outlined in Sect. 3.2.1, one can undertake the 4-stage classification problem. To ascertain if this problem is solvable, we have examined if there are genes whose (average) expression levels change monotonically with the progression of a cancer. Fortunately, numerous such genes are found, suggesting that the problem is solvable. Figure 3.2 shows three such genes, namely *LANCL3* (lanC lanti-biotic synthetase component c-like protein), *MFAP2* (microfibrillar-associated protein) and *PPA1* (pyrophosphatase).

While the average levels of these three genes each change monotonically with cancer progression, they may not necessarily represent the best genes whose

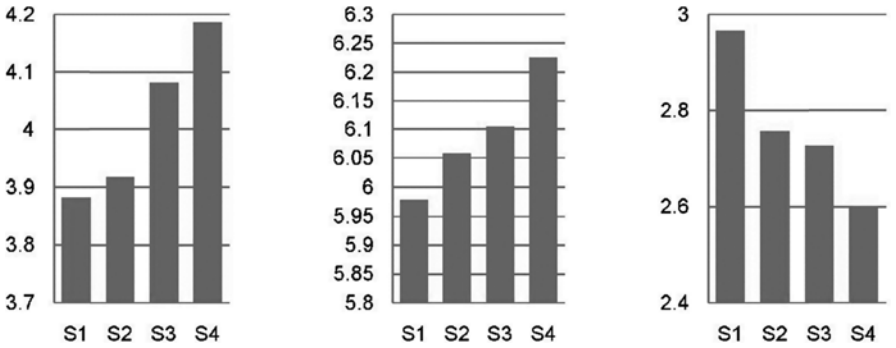

Fig. 3.2 The average gene-expression levels of three genes represented by three panels from *left* to *right*, *LANCL3*, *MFAP2* and *PPA1*, over all samples in each stage for stages S = 1, 2, 3 and 4. The y-axis is the average fold-change of gene-expression levels across all samples of a specific stage in cancer *versus* control samples, and the x-axis is the stage axis. The figure is adapted from Cui et al. (2011b)

expression levels are most informative in predicting cancer stages for individual tissue samples. To find out, an exhaustive search was made for the best K-gene discriminator, for $2 \leq K \leq 10$, for the 4-stage classification problem. The combination (*DPT, EIF1AX, FAM26D, IFITM2, LOC401498, OR2AE1, PRRG1, REEP3, RTKN2*) was found to be the best 9-gene signature for gastric cancer staging, and (*CPS1, DEFA5, DES, DMN, GFRA3, MUC17, OR9G1, REEP3, TMED6, TTN*) represents the best 10-gene marker, achieving 84.0 % and 90.0 % 4-stage classification consistencies with the pathologists who did the original staging, respectively (Cui et al. 2011b).

The following table lists the functions of these marker genes, which were retrieved from the GeneCards database (Rebhan et al. 1997), to give the reader a sense about what functional genes may serve as good markers for cancer staging. Interestingly, the two lists have very little in common with only one gene, *REEP3*, shared by the two lists plus a pair of homologous genes, *OR2AE1* and *OR9G1,* in the two lists as shown in the following table. Even by examining cellular level functions, the two sets of pathways enriched with the two gene lists have very little in common. This suggests that there is probably a sizeable set of genes whose expression patterns are informative for the determination of cancer stages, and it just happens that these two lists give rise to the two best discriminators (Table 3.1).

As in the case of cancer typing, the discrepancy between the pathologist-assigned stages and gene-expression-based staging could be due to various reasons as discussed in Sect. 3.2.1. One useful effort will be to refine both definitions through collaboration between cancer pathologists and cancer data analysts. Such a joint effort to identify reasons for staging discrepancies by the two approaches should lead to a refinement of the criteria used by both parties in an iterative fashion until there is convergence. Such an exercise could lead to improvement in cancer-staging based on gene-expression data in a systematic manner. Another important issue is that the current 4-stage classification scheme for measuring cancer progression is probably somewhat arbitrary. There is no strong evidence to support the operational premise

Table 3.1 Functional annotation of the signature genes

Gene name	Function
DPT (dermatopontin)	An extracellular matrix protein involved in cell-matrix interaction and matrix assembly
EIF1AX (ukaryotic translation initiation factor 1α)	An essential translation initiation factor
FAM26D (family with sequence similarity 26, member D)	A pore-forming subunit of a voltage gated ion channel
IFITM2 (interferon induced transmembrane protein 2)	An *IFN*-induced protein that inhibits the entry of viruses to the host cell cytoplasm
LOC401498 (a hypothetical protein)	No function has been identified
OR2AE1 (olfactory receptor 2AE1)	A hormone receptor responsible for recognition and *G* protein-mediated transduction of odorant signals
PRRG1 (proline-rich gamma-carboxyglutamic acid protein 1)	The protein containing two functional motifs generally found in signaling and cytoskeletal proteins
REEP3 (receptor accessory protein 3):	May enhance the cell-surface expression of odorant receptors
RTKN2 (rhotekin 2)	May have an important role in lymphopoiesis
CPS1 (carbamoyl-phosphate synthase):	Important in removing excess ammonia from the cell through the urea cycle
DEFA5 (defensin α5)	Has antimicrobial activity and kills microbes by permeabilizing their plasma membrane
DES (intermediate filament protein)	Forms a fibrous network connecting myofibrils to each other and to the plasma membrane
DMN (dystrophin)	A cohesive protein linking actin filaments to another support protein that resides on the inside surface of each muscle fiber's plasma membrane
GFRA3 (glial cell-derived neurotrophic factor family receptor)	Mediates the artemin-induced autophosphorylation and activation of the RET (rearranged during transfection) receptor tyrosine kinase
MUC17 (cell surface associated mucin 17)	Active in maintaining homeostasis on mucosal surfaces
OR9G1 (olfactory receptor, family 9)	May serve as a hormone receptor like *OR2AE1* in the above
TMED6 (transmembrane emp24 protein transport domain)	A *HNF1*α (hepatic nuclear factor 1α) regulated transporter
TTN (connectin)	Contributes to the balance of forces between the two halves of the sarcomere by providing connections at the level of individual microfilaments

that the development of a cancer has four distinct phases, but not three or five or even a continuous progression without obvious phases and phase transitions, say, in terms of their probabilities to metastasize. To rigorously address this issue computationally, it will require not only transcriptomic data of cancer *versus* control tissues, but also data regarding metastases. This is clearly an area where computational approaches could assist in making fundamental and highly meaningful advances.

3.2.3 Cancer Grading

Cancer grading is a less developed area compared to cancer typing and staging. Only a handful of grading systems have been proposed for some cancer types since Bloom and Richardson developed the first grading system for breast cancer in 1957 (Bloom and Richardson 1957). Similar classifications include the Gleason system for prostate cancer (Gleason 1966; Gleason and Mellinger 1974), the Fuhrman method for kidney cancer (Fuhrman et al. 1982) and the approach proposed by Goseki et al. for gastric cancer (Goseki et al. 1992). As of now, only a few grading systems have been developed based on molecular information, such as the Nottingham grading system for breast cancer (Simpson et al. 2000) and the work by Cui et al. for gastric cancer (Cui et al. 2011b). The main challenge here is that, unlike cancer typing and staging, for which some molecular level information has already been used, cancer grading has been solely based on morphologic data of cancer cells and decided by cancer pathologists. Hence, there may be a large gap between pathologist-assigned grades and molecular-level commonalities among samples of the same grade. An example is given here to illustrate the possibility of using transcriptomic data to grade cancer tissues and point out possible issues with the existing grading procedures.

We continue to use the same gastric cancer dataset introduced in Sect. 3.2.1. Out of the 80 gastric cancer tissues, 54 have grades assigned by cancer pathologists (Cui et al. 2011b), so only these data are used for developing a computational method for grading a tumor based on its gene-expression data. Of the 54 tissues, 8 are well differentiated (WD), 9 moderately differentiated (MD), 35 poorly differentiated (PD) and 2 undifferentiated (UD), with the patients' data given in Table 3.2. The aim here is to identify a set of genes whose expression patterns can well distinguish among the four grades of gastric cancer.

As in cancer staging, one can determine if some genes have expression levels that change monotonically with change in cancer grades from highly differentiated to undifferentiated. Using this criterion, 99 such genes were found. For each of these genes, its average fold-change among samples of each grade exhibits a monotonic relationship with the grade list WD-MD-PD-UD from the least malignant to the most malignant, suggesting that the current grading scheme for gastric cancer does have some molecular basis. These genes include *POF1B* (premature ovarian failure 1β), *MET* (hepatocyte growth factor receptor), *CEACAM6* (carcinoembryonic antigen-related cell adhesion molecule), *ZNF367* (zinc finger protein involved in transcriptional activation of erythroid genes), *GKN1* (gastrokine-1 with strong anticancer activity), *LIPF* (gastric lipase with lipid binding and retinyl-palmitate esterase activity), *SLC5A5* (a glutamate transporter), *MUC13* (cell surface associated mucin), *CLDN1* (senescence-associated epithelial membrane protein), *MMP7* (matrix metalloproteinase) and *ATP4A* (ATPase, H+/K + transporting, α). Figure 3.3 shows four examples of these genes in terms of their averaged expression levels *versus* cancer grades across samples of each cancer grade.

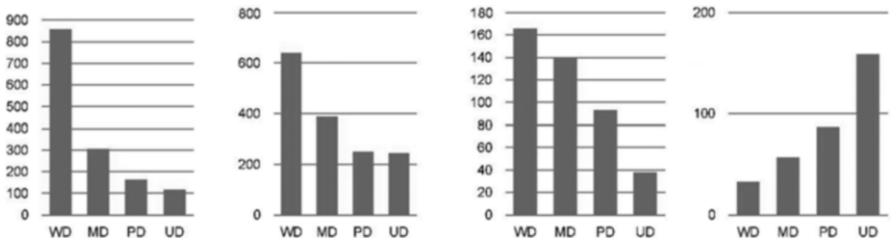

Fig. 3.3 The average gene-expression levels of four genes, *CEACAM6, MUC13, CLDN1* and *PGA4*, over gastric cancer samples of each grade for grades WD, MD, PD and UD. The definitions of the y- and x-axis are the same as in Fig. 3.2. Adapted from Cui et al. (2011b)

Intuitively one may expect that some combinations of the 99 genes should give a good classification among the four grades. However, this may not necessarily be the case for the same reason as discussed in Sect. 3.2.2. Instead, a 19-gene combination is identified, whose expression fold-changes gave a 79.2 % classification consistency with pathologist-assigned grades on two combined grades, namely "highly differentiated" covering the WD and MD samples and "poorly differentiated" for the PD and UD samples, using the algorithm of Sect. 3.2.1. It takes a minimum of 198 genes to give a 4-grade classification at a comparable classification consistency, specifically at 74.2 %.

There may be multiple reasons for the relatively low consistency levels between the pathologist-decided and gene-expression-based grading results, but one key reason, we suspect, may be that the morphological information-based grade arrived at by pathologists may not be as informative in terms of their prognostic values as it could be, at least not on this dataset, indicating the possible limitations of the current approaches and a need for improved techniques.

3.3 Discovering (Sub)Types, (Sub)Stages and (Sub)Grades Through Data Clustering

The analysis presented in Sect. 3.2 is based on the assumption that the pathologist-assigned cancer types, stages and grades are generally correct, i.e., they reflect, to a large extent, the true molecular level commonalities of cancer samples within each type (or stage, grade) and differences across cancer samples of different types (or stages, grades). A more general cancer typing (or staging, grading) problem is to identify cancer types (or stages, grades) when the information of human-designated types (stages and grades) is not available. The question addressed here is: *Can one possibly discover types or subtypes of a cancer based on the similarities among expression patterns of some (to-be-identified) genes* among a subset of cancer and matching control samples. To put it in a more specific context: when given a collection of gene-expression data collected on leukemia samples consisting of four types

of leukemia, namely ALL, AML, CLL and CML, but without any labels, *is it possible to rediscover the four types of leukemia from the given samples based solely on their gene-expression data*? The answer is: Yes, but it may take a lot of computing time.

From a computational perspective, this represents a different type of data analysis problem from those discussed in Sect. 3.2, which are called *classification* problems. The main issue there was: *Given a set of objects, each labeled to belong to a specific class, can one identify "features" that can accurately predict the class label (*e.g.*, stages or types) of each object based on the features*? For the current problem, the question is: *For the same set of objects, can one partition all the objects into a few classes so that objects in each class share some common features that are not shared by objects in other classes*? Using computer science terminology, this is a *clustering* problem.

Clustering techniques have long been used in gene-expression data analyses (Ben-Dor et al. 1999; Wu et al. 2004; D'haeseleer 2005). Through identification of sample groups sharing similar expression patterns of some genes, researchers have identified various previously unknown subclasses of human diseases. The earliest work in cancer class discovery based on gene-expression data was published by Golub et al., which showed that without prior knowledge, the algorithm "discovered" two subtypes of leukemia, namely, AML and ALL, based on the distinct gene-expression patterns among samples of the two subtypes (Golub et al. 1999). Other discoveries of cancer subtypes include: (1) the discovery of five subtypes of breast cancers based on gene-expression patterns, namely, luminal A, luminal B, basal-like, normal-like and *ERBB2*+ groups, which were found to have clinical implications (Livasy et al. 2006); (2) a recent study that classifies colon cancer into six subtypes based on distinct genomic mutation patterns in the samples, namely samples with or without *BRAF*, *KRAS* and *P53* mutations, CpG island methylation patterns, DNA mismatch repair status and the chromosomal instability level. The study also showed clinical relevance of the six subtypes (Marisa et al. 2013); and (3) a study that showed improvement in subtyping over the previously determined subtypes of leukemia using gene-expression data (Yeoh et al. 2002).

These examples signify the importance that the to-be-discovered new subtypes must have clinical relevance. Otherwise such an analysis may lead to clustering results that group cancer samples according to their growth rates, which may share similar expression patterns of some genes but not any common driving or facilitating mechanisms in cancer development, hence limiting their usefulness from a clinical perspective.

Recent studies have revealed one key inadequacy in the current clustering techniques in discovering subgroups having common or similar gene-expression patterns, which are distinct from other subgroups. Specifically, a major issue is that the clustering techniques require a pre-defined subset of genes, based on which tissue samples are grouped according to the similarities in expression patterns of these genes. This, however, is too restrictive for discovering novel subgroups that may have similar expression patterns of some genes that cannot be determined in advance. The computational difficulty in handling this more general clustering problem is that for a problem with m differentially expressed genes, 2^m combinations of genes

need to be considered in order to identify a subset of the *m* genes sharing similar expression patterns among some samples. When *m* is relatively large, say even in the range of a few tens, this clustering problem becomes computationally intractable. A more powerful clustering strategy is needed to solve such problems, and *bi-clustering* is one such technique (Van Mechelen et al. 2004).

To understand the basic principles of a bi-clustering algorithm, one can represent a gene-expression dataset as a numeric matrix with each row representing a gene, each column representing a paired (cancer *versus* control) sample, and each entry in the matrix having the log-ratio value between the expression levels of the corresponding gene in the corresponding sample pair. Two genes are considered to have similar expression patterns for a subset of samples if the correlational coefficient between the two genes-corresponding rows across the samples-corresponding columns is above some defined threshold. A *bi-clustering* problem is defined as that locating all (maximal) sub-matrices, in each of which the correlational coefficient between each pair of rows across the samples defined by the sub-matrix is above the specified threshold. Each so defined sub-matrix is called a *bi-cluster*. Clearly, a bi-clustering problem is substantially more general than the traditional clustering problem, in that it enables one to discover previously unknown subclasses of a cancer class (e.g., type, stage or grade). The generality of a bi-clustering problem also makes it considerably more difficult to solve computationally.

A number of algorithms have been proposed to solve this challenging problem (Madeira and Oliveira 2004; Van Mechelen et al. 2004). To assess the effectiveness of the bi-clustering approach in subgroup discovery, we have applied QUBIC (Li et al. 2009), a bi-clustering method we previously developed, to gene-expression data of three leukemia types, ALL, MLL and AML, mixed together with their type information removed. The algorithm can accurately recover the three subtypes of leukemia as shown in Fig. 3.4, suggesting the general feasibility in discovering subtypes from gene-expression data of multiple samples of the same cancer type.

This technique has also been applied to the 80 pairs of gastric cancer expression data for the discovery of possible subgroups among the samples, which led to the identification of 20-plus bi-clusters. Some of these bi-clusters represent previously uncharacterized subtypes of gastric cancer. For example, Fig. 3.5 shows one bi-cluster defined by 42 genes, for which the 80 samples fall into two groups, each sharing common expression patterns of the 42 genes but different between the two groups, specifically the light-gray subset on the left and the dark-gray subset on the right in the figure. Further analyses suggest that the two subgroups may belong to two known subtypes of gastric cancer, namely intestinal and diffuse subtypes (Shah et al. 2011). This conclusion is based on the observation that six of the 42 genes, namely *CNN1, MYH11, LMOD1, MAOB, HSPB8* and *FHL1*, have previously been reported to be differentially expressed between the intestinal and the diffuse subtypes of gastric cancer, which all show similar expression patterns among samples in each subgroup in the figure.

Such a bi-clustering analysis can also be used for discovery of cancer stages and grades. The approach is to first identify genes whose expression patterns change with alterations in stage or grade and then conduct bi-clustering analyses using such genes as the gene set like the above analysis on cancer subtypes.

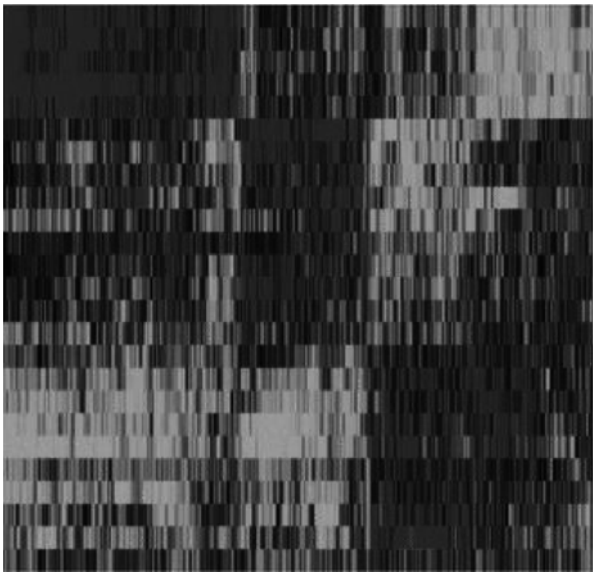

Fig. 3.4 An illustration of the identified three subtypes of leukemia based on gene-expression data using the bi-clustering method QUBIC without using *a prior* knowledge about the three subtypes. The *rows* and *columns* represent genes and samples, respectively, and *dark gray* and *light gray* represent up- and down-regulations, respectively

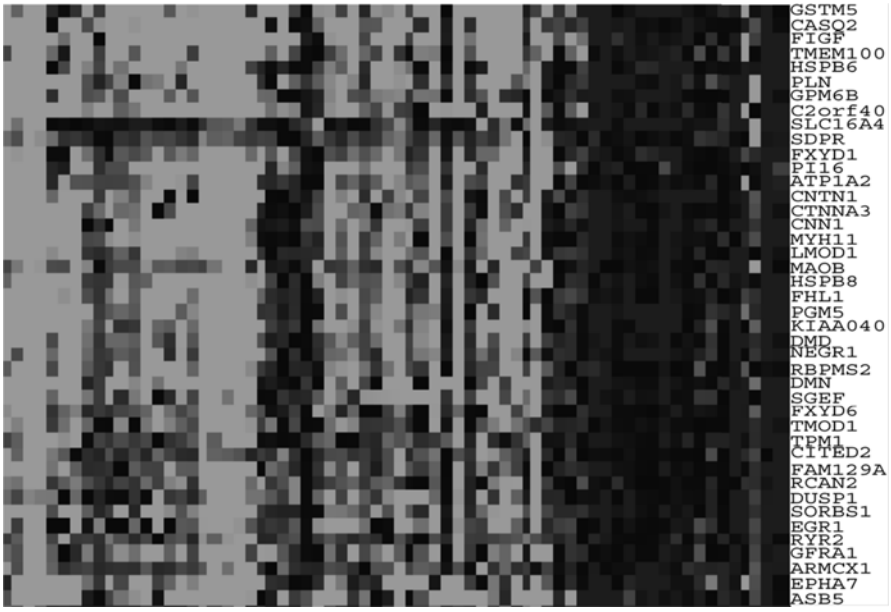

Fig. 3.5 A bi-clustering result based on 42 genes (*listed along the right side* of the figure) and 80 paired samples (*columns*). The patterns suggest that the 80 patients fall into two subtypes, intestinal and diffuse subtypes. Adapted from Cui et al. (2011a)

3.4 Challenging Issues

The availability of genome-scale transcriptomic data for a variety of cancer samples has enabled molecular information-based typing, staging and grading on more objective and scientific grounds. Along with this opportunity also comes a number of challenging technical issues in dealing with the complexity of the data and discovering samples sharing distinct gene-expression patterns with statistical significance. A few such challenges that must be addressed in order to make cancer typing, staging or grading analyses done in an informative and reliable manner are listed below.

3.4.1 Identification of Pathway-Level Versus Gene-Level Signatures

The basic premise for cancer typing (and similarly staging, grading) using classification or clustering techniques is that some genes exhibit similar expression patterns in cancer samples of the same type, which are not shared by cancers in other types. While this is probably true for some genes and cancers as shown in this chapter, there is no reason to believe that this has to be true universally. The reason is that cancers sharing certain phenotypic characteristics may tend to behave similarly at the biochemical pathway level rather than at the individual gene level. For example, the repression of the apoptosis system could be accomplished through functional state changes in numerous different ways such as the inhibition of *P53* transcription, *P53* gene mutations, over-expression of various survival pathways, the activation of anti-apoptotic members of the *BCL2* family, and over-expression of certain oncogenes. There are even multiple ways to repress apoptosis just through different ways of inhibiting the function of *P53*, such as repression of *P53*'s expression transcriptionally or epigenomically, over-expression of its inhibitory binding partner *MDM2*, prevention of the *P53* protein from entering the nucleus or inhibition of *P53*'s function through posttranslational modification (see Chap. 7 for details). Hence, an improved strategy for gene-expression-based cancer typing needs to take this fact into consideration. An improved strategy may need to first identify *equivalent* gene groups, each defined as genes whose expression changes may lead to the same effects at the pathway level. The challenge is how to identify such equivalent gene groups, which, we believe, requires novel ideas knowing that the current understanding of cancer-relevant pathways is far from complete.

3.4.2 *Close Collaboration Between Data Analysts and Pathologists May Be Essential*

Another challenge in using computational techniques for cancer typing (or staging, grading) lies in how to optimally integrate the experience of cancer pathologists in defining cancer types and the molecular information hidden in the *omic* data. A common practice, as shown above, has been to statistically link cancer samples, defined as the same type by pathologists, to a set of genes with common expression patterns, which are distinct from cancer samples of the other types. An issue encountered with such an approach is what to do next when the computational methods give rise to staging results different from those by pathologists, knowing that both approaches could have errors. An important message to convey here is that it is essential for cancer pathologists and *omic* data analysts to collaborate in order to resolve inconsistent results, and better yet to develop general protocols for mapping the knowledge of onco-pathologists to computer-based cancer typing, staging and grading procedures.

3.4.3 *Capturing Complex Relationships Among Gene-Expression Patterns*

Another challenging issue is to identify complex relationships among gene expression data. For example, some cellular regulation may be triggered when the difference between the concentrations of certain gene products exceed a certain range, rather than their actual expression levels increasing above some threshold. Oxidative stress, defined as the difference between the abundance of oxidant molecules (such as ROS) and that of antioxidants (see Chap. 8 for details), serves as a good example here. Specifically it is the difference between the abundances of ROS molecules and the antioxidant species, rather than the abundance of one individual molecular species like ROS, that triggers oxidative-stress responses when it is beyond some threshold. Basically more general models are needed for capturing the complex relationships among gene expression data than simply up-or-down expression levels. The problem here is to detect non-trivial mathematical relationships among some genes, which are shared by some subgroup of samples. Clearly this represents a substantially more complex problem in identifying genes similar expression patterns, which, if solvable, can help to solve substantially more complex clustering problems.

3.5 Concluding Remarks

The state of the art in cancer typing, staging and grading relies heavily on morphological information of cancer cells, along with limited molecular level data. The limitation of such approaches is obvious since they are not connected with the

detailed molecular mechanism(s), raising an urgent need for improved cancer characterization using *omic* data. The importance in moving in this direction is clear, as knowing that typing, staging and grading have important implications to prognosis as well as selection of the optimum treatment plan(s). Large scale *omic* data, such as transcriptomic data, probably contain all or the majority of the information about the underlying cancer in terms of its driving force, growth mechanism and ability to invade and metastasize. By linking such information to typing, staging and grading, one can potentially develop more effective ways to assess the level of development and malignancy of a cancer. To render *omic* data-based cancer typing, staging and grading prediction impactful, collaboration between cancer pathologists and *omic* data analysts is the key.

There are two types of computational techniques that can assist in cancer typing, staging and grading. One relies on training datasets in which cancer samples are labeled with specific types, stages and grades by cancer pathologists; the problem is to extend this knowledge to enable computer programs to make the same calls by identifying genes whose expression patterns correlate well with the specified types, stages or grades. This is an example of what is termed a classification problem, or *supervised learning* as referred to in the field of data mining. The other does not require a training dataset; instead the problem is to determine if a given group of cancer samples can be partitioned into subgroups so that each shares common expression patterns among some to-be-identified genes, but distinct from other cancer samples. This approach is denoted as a clustering problem, or an *un-supervised learning* problem. Various challenging computational problems exist that await improved techniques, thus making computer-based decisions substantially more reliable than the state-of-the-art, including: (1) going beyond the simple similarity measures between gene expression to capture more complex relationships among gene-expression data of different cancer samples of the same type, stage or grade; and (2) more integrated approaches to cancer typing, staging and grading through a refinement of the existing classification schemes involving feedback from pathologists and computational prediction.

Appendix

Table 3.2 Patient data used in the analysis in Sect. 3.2

Patient ID	Age	Gender	Histologic type	Grade	Stage	Smoking	Alcohol	Weight
1	54	F	WMD	G2	III	0	0	70
2	62	F	WMD	G1	IIIA	0	0	60
3	53	M	WMD	G2	IIIB	0	0	60
4	51	M	WMD	G2	IIIB	1	0	–
5	73	M	WMD	–	IB	0	0	63
6	41	M	WMD	G2	II	–	–	–
7	59	M	WMD	G1	III	1	1	51

(continued)

Table 3.2 (continued)

Patient ID	Age	Gender	Histologic type	Grade	Stage	Smoking	Alcohol	Weight
8	68	M	WMD	G2	IV	0	0	48
9	56	F	WMD	G1	IIIA	0	0	45
10	43	F	WMD	G1	III	0	0	55
11	71	F	WMD	G2	III	0	0	42
12	65	M	WMD	G2	IIIA	0	0	70
13	55	M	WMD	G2	III	0	0	69
14	55	M	WMD	G2	IIIB	0	0	74
15	62	F	WMD	G1	IV	–	–	–
16	41	F	SRC	–	IV	0	0	43
17	42	M	SRC	–	III	0	0	60
18	68	M	SRC	–	III	0	0	50
19	50	M	SRC	–	III	0	0	62
20	55	M	SRC	–	III	0	0	50
21	34	M	SRC	–	III	0	0	90
22	63	M	PD	G3	IIIB	1	1	–
23	56	M	PD	G3	IIIB	1	1	–
24	71	M	PD	G3	IIIB	1	0	–
25	55	F	PD	G3	IIIB	0	0	63
26	64	M	PD	G3	IIIB	0	0	55
27	53	F	PD	G3	IIIB	0	0	77
28	56	M	PD	G3	IIIB	1	0	55
29	53	M	PD	G2–G3	III	0	0	62
30	71	M	PD	G3	III	0	0	60
31	58	M	PD	G2–G3	III	0	0	50
32	42	M	PD	G3	IB	0	0	52
33	65	F	PD	G3	IIIA	0	0	–
34	50	M	PD	G3	III	1	0	47
35	59	M	PD	G3	III	0	0	57
36	75	M	PD	G3	III	0	0	65
37	40	M	PD	G3	III	0	1	80
38	51	F	PD	G3	III	1	0	52
39	67	F	PD	G3	IV	0	0	48
40	65	F	PD	G3	IIIA	0	0	53
41	53	F	PD	G3	IIIA	1	0	60
42	60	F	PD	G3	IIIB	0	0	60
43	70	M	PD	G3	II	1	0	59
44	56	F	PD	G3	II	0	0	74
45	78	F	PD	G3	IIIB	0	0	39
46	65	M	PD	G3	III	0	1	70
47	68	M	PD	G3	III	1	1	69

(continued)

Table 3.2 (continued)

Patient ID	Age	Gender	Histologic type	Grade	Stage	Smoking	Alcohol	Weight
48	57	F	PD	G3	IIIA	0	0	61
49	68	F	PD	G3	III	–	–	–
50	61	M	PD	G2–G3	III	1	0	70
51	55	M	PD	G3	III	–	–	–
52	67	F	PD	G3	II	–	–	–
53	50	F	PD	G3	III	–	–	–
54	62	F	MC	–	III	0	0	70
55	55	M	MC	–	IIIB	0	0	60
56	57	M	MC	G2	IIIA		–	65
57	74	M	MC	–	IB	0	0	62
58	58	M	MC	G3	IV	0	0	66
59	76	M	MC	–	II	0	0	70
60	54	M	MC	–	III	1	1	49
61	47	M	(tublar)	–	IB	1	1	65
62	49	M	(tubular/ papillary)	–	III	1	1	60
63	76	F	(undifferentiated)	G4	II	0	0	–
64	51	M	(undifferentiated)	G4	II	–	NA	70
65	69	F	(squamous cell)	–	III	0	0	50
66	65	M	(squamous cell)	G3	III	0	1	50
67	36	M	(ulcerative)	G3	IIIA	1	0	60
68	75	F	(ulcerative)	G2–G3	IV		–	40
69	69	M	(mucous cell type)	G3–G4	III	0	0	55
70	81	M	(adenosquamous)	–	III	1	0	56

References

Albain KS, Barlow WE, Shak S et al. (2010) Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. Lancet Oncol 11: 55-65

Arranz EE, Vara JA, Gamez-Pozo A et al. (2012) Gene signatures in breast cancer: current and future uses. Transl Oncol 5: 398-403

Ben-Dor A, Shamir R, Yakhini Z (1999) Clustering gene expression patterns. Journal of computational biology : a journal of computational molecular cell biology 6: 281-297

Beutler E (2001) The treatment of acute leukemia: past, present, and future. Leukemia 15: 658-661

Bloom HJ, Richardson WW (1957) Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer 11: 359-377

Chen JJ, Knudsen S, Mazin W et al. (2012) A 71-gene signature of TRAIL sensitivity in cancer cells. Mol Cancer Ther 11: 34-44

Cho SH, Jeon J, Kim SI (2012) Personalized medicine in breast cancer: a systematic review. J Breast Cancer 15: 265-272

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20: 273-297

Cui J, Chen Y, Chou WC et al. (2011a) An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer. Nucleic acids research 39: 1197-1207

Cui J, Li F, Wang G et al. (2011b) Gene-expression signatures can distinguish gastric cancer grades and stages. PLoS One 6: e17819

D'haeseleer P (2005) How does gene expression clustering work? Nature Biotechnology 23: 1499-1501

Duan KB, Keerthi SS (2005) Which is the best multiclass SVM method? An empirical study. Multiple Classifier Systems 3541: 278-285

Eddy JA, Sung J, Geman D et al. (2010) Relative expression analysis for molecular cancer diagnosis and prognosis. Technol Cancer Res Treat 9: 149-159

Erten S, Chowdhury SA, Guan X et al. (2012) Identifying stage-specific protein subnetworks for colorectal cancer. BMC Proc 6 Suppl 7: S1

Fuhrman SA, Lasky LC, Limas C (1982) Prognostic significance of morphologic parameters in renal cell carcinoma. Am J Surg Pathol 6: 655-663

Gleason DF (1966) Classification of prostatic carcinomas. Cancer Chemother Rep 50: 125-128

Gleason DF, Mellinger GT (1974) Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J Urol 111: 58-64

Golub TR, Slonim DK, Tamayo P et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537

Goodison S, Sun Y, Urquidi V (2010) Derivation of cancer diagnostic and prognostic signatures from gene expression data. Bioanalysis 2: 855-862

Goseki N, Takizawa T, Koike M (1992) Differences in the mode of the extension of gastric cancer classified by histological type: new histological classification of gastric carcinoma. Gut 33: 606-612

Guyon I, Weston J, Barnhill S et al. (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46: 389-422

Inza I, Larranaga P, Blanco R et al. (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med 31: 91-103

Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8: 118-127

Kim WJ, Kim SK, Jeong P et al. (2011) A four-gene signature predicts disease progression in muscle invasive bladder cancer. Mol Med 17: 478-485

Lauren P (1965) The Two Histological Main Types of Gastric Carcinoma: Diffuse and So-Called Intestinal-Type Carcinoma. An Attempt at a Histo-Clinical Classification. Acta pathologica et microbiologica Scandinavica 64: 31-49

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

Liong ML, Lim CR, Yang H et al. (2012) Blood-based biomarkers of aggressive prostate cancer. PLoS One 7: e45802

Livasy CA, Karaca G, Nanda R et al. (2006) Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 19: 264-271

Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform 1: 24-45

Marisa L, de Reynies A, Duval A et al. (2013) Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med 10: e1001453

Mroz EA, Rocco JW (2012) Gene expression analysis as a tool in early-stage oral cancer management. J Clin Oncol 30: 4053-4055

Mukherjee S (2010) The emperor of all maladies: a biography of cancer. Scribner,

Ramaswamy S, Tamayo P, Rifkin R et al. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences of the United States of America 98: 15149-15154

Rebhan M, ChalifaCaspi V, Prilusky J et al. (1997) GeneCards: Integrating information about genes, proteins and diseases. Trends Genet 13: 163-163

Reis-Filho JS, Pusztai L (2011) Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet 378: 1812-1823

Rodenhiser DI, Andrews JD, Vandenberg TA et al. (2011) Gene signatures of breast cancer progression and metastasis. Breast Cancer Res 13: 201

Shah MA, Khanin R, Tang L et al. (2011) Molecular classification of gastric cancer: a new paradigm. Clin Cancer Res 17: 2693-2701

Shedden K, Taylor JM, Enkemann SA et al. (2008) Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med 14: 822-827

Simpson JF, Gray R, Dressler LG et al. (2000) Prognostic value of histologic grade and proliferative activity in axillary node-positive breast cancer: results from the Eastern Cooperative Oncology Group Companion Study, EST 4189. J Clin Oncol 18: 2059-2069

Slodkowska EA, Ross JS (2009) MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients. Expert Rev Mol Diagn 9: 417-422

Starmans MH, Krishnapuram B, Steck H et al. (2008) Robust prognostic value of a knowledge-based proliferation signature across large patient microarray studies spanning different cancer types. Br J Cancer 99: 1884-1890

Stewart BW, Kleihues P (2003) World cancer report. IARC Press,

Tibshirani R, Hastie T, Narasimhan B et al. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences of the United States of America 99: 6567-6572

Van Mechelen I, Bock HH, De Boeck P (2004) Two-mode clustering methods: a structured overview. Stat Methods Med Res 13: 363-394

Warburg O (1956) On the origin of cancer cells. Science 123: 309-314

Weigelt B, Baehner FL, Reis-Filho JS (2010) The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. The Journal of pathology 220: 263-280

Wu S, Liew AW, Yan H et al. (2004) Cluster analysis of gene expression data based on self-splitting and merging competitive learning. IEEE Trans Inf Technol Biomed 8: 5-15

Yamagiwa K, Ichikawa K (1918) Experimental study of the pathogenesis of carcinoma. The Journal of Cancer Research 3: 1-29

Yeoh EJ, Ross ME, Shurtleff SA et al. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1: 133-143