Published in final edited form as:

IEEE Rev Biomed Eng. 2012 ; 5: 74–87. doi:10.1109/RBME.2012.2212427.

Multiscale Integration of -Omic, Imaging, and Clinical Data in Biomedical Informatics

John H. Phan¹ [Member, IEEE], Chang F. Quo¹, Chihwen Cheng², and May Dongmei Wang³ [Senior Member, IEEE]

¹Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA

²Electrical and Computer Engineering Department, Georgia Institute of Technology, Atlanta, GA 30332 USA

³Biomedical Engineering Department, Electrical and Computer Engineering, Winship Cancer Institute, Parker H. Petit Institute of Bioengineering and Biosciences, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA

Abstract

This paper reviews challenges and opportunities in multiscale data integration for biomedical informatics. Biomedical data can come from different biological origins, data acquisition technologies, and clinical applications. Integrating such data across multiple scales (e.g., molecular, cellular/tissue, and patient) can lead to more informed decisions for personalized, predictive, and preventive medicine. However, data heterogeneity, community standards in data acquisition, and computational complexity are big challenges for such decision making. This review describes genomic and proteomic (i.e., molecular), histopathological imaging (i.e., cellular/tissue), and clinical (i.e., patient) data; it includes case studies for single-scale (e.g., combining genomic or histopathological image data), multiscale (e.g., combining histopathological image and clinical data), and multiscale and multiplatform (e.g., the Human Protein Atlas and The Cancer Genome Atlas) data integration. Numerous opportunities exist in biomedical informatics research focusing on integration of multiscale and multiplatform data.

I INTRODUCTION

BIOMEDICAL informatics is essential for personalized, predictive, and preventive medicine. Each of these medical applications requires data from multiple sources and multiple scales. Fig. 1 illustrates three scales of biomedical data: 1) molecular (e.g., genomic and proteomic, or -omic); 2) cellular/tissue (e.g., histopathological imaging); and 3) patient (e.g., clinical). In addition, databases of biological knowledge, primarily in the form of literature, are used for integrative analysis. Thus, multiscale data integration is a key component of biomedical informatics. The drive to realize personalized, predictive, and preventive medicine provides numerous research opportunities in multiscale data integration for both the biomedical and clinical communities. For example, patient data from a variety of modalities can help to determine the most successful course of treatment on a personalized basis. Molecular and cellular/tissue data can be used to predict a patient's prognosis given a particular treatment course. Molecular profiles as well as patient history

may help in preventing the onset of disease for high-risk patients. This paper presents case studies focusing on integration of -omic, imaging, and clinical data to illustrate some of the benefits and challenges related to multiscale biomedical data integration. These case studies represent only a subset of larger integrative studies such as the Virtual Physiological Human (VPH) Physiome Project [1].

Section II summarizes four scales of biomedical data, describing acquisition techniques and clinical applications. Section III presents five case studies of multiscale biomedical data integration. These case studies include: 1) genomic data meta-analysis; 2) histopathological image data normalization; 3) knowledge-driven bioinformatics; 4) the Human Protein Atlas (HPA); and 5) the Cancer Genome Atlas (TCGA). Although these case studies represent only a few of the numerous biomedical applications, they are intended to highlight the potential benefits and current challenges of state-of-the-art biomedical data integration research.

II BIOMEDICAL DATA

Biomedical data is acquired from a variety of sources and diversity in procedure. For instance, it may be gathered: 1) directly, using various experimental protocols and equipment in the laboratory and clinic; 2) indirectly, through increasingly collaborative partnerships with affiliated domain specialists; and 3) remotely, from groups that share their data publicly. Each of these avenues customizes the biomedical data gathered to address specific interests and needs that are rarely in complete agreement with the others. As a result, the acquired data is unstructured—i.e., it is documented in potentially very different formats—and is not guaranteed to be directly comparable to similar data gathered from other laboratories. As a result, methods for handling increasingly heterogeneous data have become a prominent focus of biomedical informatics. In order to effectively use heterogeneous biomedical data for personalized, predictive, and preventive medical applications, it is important to understand the biological and technological factors for heterogeneity. This section summarizes several biomedical data types and acquisition methods on multiple scales, including molecular, cellular/tissue, and patient. In addition, it describes some biomedical knowledge bases that are increasingly used in integrative bioinformatics.

A. Molecular Level Data

Molecular level biomedical data is the quantitative representation of the activities and properties of genes, proteins, and other biomolecules that are part of biological systems. The benefits of molecular level data for biomedical informatics are significant. For example, gene expression-based predictors can potentially improve the accuracy of early disease detection [2], [3], [4]. Thus, studies of gene expression that aim to identify genomic biomarkers of cancer are a major component of bioinformatics research [5]. Rapidly evolving high-throughput (i.e., capable of assaying thousands of genes simultaneously) microarray and next-generation sequencing (NGS) technologies have produced a large amount of gene expression data [6], [7]. Numerous review papers exist covering general concepts and applications of both microarray [8], [9], [10] and NGS technology [11], [12]. In addition to gene expression, microarray and NGS technologies are capable of quantifying

other genomic properties including single nucleotide polymorphisms (SNPs) [13], [14], DNA copy number [15], [16], DNA methylation [17], [18], and microRNA activity [19], [20], among others.

Gene expression profiles, and other genomic properties that affect gene expression, are related to phenotype. However, this relation is indirect, and gene expression may also be influenced by environmental factors (e.g., metabolism or microenvironment). Moreover, the abundance of proteins, molecules that largely contribute to biological structure and function, is not necessarily correlated with gene expression [21]. Thus, quantification of protein abundance and activity is important. Although protein analysis is much more difficult than genomic analysis because of biochemical factors, technologies such as mass spectrometry [22], [23] and protein microarrays [24], [25] are promising tools for high-throughput quantification of proteins.

Bioinformaticians often resort to meta-analysis, i.e., the combination or analysis of multiple molecular datasets in order to increase sample size. Increasing sample size improves statistical confidence, which is important in applications such as identification of differentially expressed genes. In this application, high statistical confidence implies that there is a low probability that uninformative genes would appear to be differentially expressed due to random chance. Unfortunately, batch effects are a major obstacle in cross-platform and cross-laboratory meta-analysis due to the diversity of technologies and protocols for acquiring molecular data. In terms of biomedical data, batch effects are differences among datasets or groups of samples due to technological factors (e.g., data acquisition technology or experimental protocol) rather than to biological factors. The case study in Section III-A discusses some challenges of genomic data meta-analysis and highlights some promising solutions, one of which is knowledge integration (Section III-C). Integration of molecular level data with cellular/tissue level data (Section III-D, HPA and Section III-E, TCGA) also enables unique clinical applications.

B. Cellular and Tissue Level Data

Molecular profiling (i.e., quantification of genomic or proteomic properties using gene or protein microarrays, NGS, or mass spectrometry) has been successfully used for applications such as cancer grading. However, molecular profiling discards spatial, morphological information in cells and tissues that can be important. For example, cancer grade is often associated with specific morphological characteristics of cell nuclei [26], [27], [28]. Evidence suggests that combined analysis of molecular profiling and morphological quantification may lead to improved clinical stratification of cancers into grades and subtypes [29]. Thus, in addition to molecular profiling technologies, it is important to understand technologies that enable morphological analysis of cells and tissues.

Technologies for morphological analysis of cells and tissues include contrast-enhanced microscopic imaging and imaging mass spectrometry (IMS). Contrast-enhancing agents for morphological tissue analysis differ in specificity and optical properties. Some staining methods, such as hematoxylin and eosin, enhance general structures such as cell nuclei and cytoplasm. Immunohistochemistry (IHC) is more specific in that it enhances antigens (or proteins) by introducing complementary antibodies attached to color-producing enzymes or

fluorophores [i.e., immunofluorescence (IF)]. Fluorophores are typically organic compounds with specific color properties. Quantum dots, semiconductor-based fluorophores, have advantages over organic fluorophores such as brightness, stability, and a wide range of tunable colors [30]. Thus, quantum dots are the focus of research for biological applications such as *in vitro* and *in vivo* tissue imaging [31], [32], cell and bio-molecular tracking [33], and intraoperative tumor detection [34]. Alternatively, IMS quantifies the molecular content of tissues while preserving the spatial properties of the sample [35]. Unlike contrast-enhanced imaging, IMS is able to obtain comprehensive molecular profiles of each spot on a tissue sample and, thus, can be used to identify novel molecular content [36].

Both single- and multiscale integrative analysis of cellular and tissue level data are clinically and biologically informative. Analysis of cellular/tissue imaging data often suffers from data dimensionality problems similar to that of molecular level data. Moreover, because of the diversity of cellular/tissue image acquisition technologies and protocols (e.g., microscopes and staining methods), diagnostic systems must be able to handle batch effects. The case study in Section III-B discusses some of the challenges and solutions in histopathological image data meta-analysis. Because cellular and tissue level data is closely related to molecular level data, large repositories of histopathological image data such as the Human Protein Atlas (HPA) [37] and the Cancer Genome Atlas (TCGA) [38], [39] often contain matching molecular data. Sections III-D and III-E describe these repositories, respectively, and summarize some of the integrative molecular and cellular/tissue research that use these repositories.

C. Patient Level Data

The National Cancer Institute broadly defines clinical data as “data obtained through patient examination or treatment” (<http://nciterns.nci.nih.gov/>). In this paper, “patient level data” refers to a data scale while “clinical data” refers to a data type on the patient level. More specifically, clinical data refers to data that describes: 1) the status of a patient; 2) parameters of therapeutic procedures; and 3) therapeutic outcome of the patient. A major area of biomedical research focuses on standardization and implementation of electronic health records (EHRs) for clinical data acquisition that can improve the efficiency and quality of healthcare. Because paper-based records are still the most widely used medium for recording patient history in the United States, parsing and conversion of handwritten records to electronic records remains an important and difficult problem [40]. In addition to challenges in the initial adoption of EHRs, hospitals and clinics tend to adopt varying functionalities of EHRs (as defined by Jha *et al.* [41]) despite the existence of broad health informatics communications standards such as Health Level Seven (HL7) [42]. Thus, due to challenges in clinical data reporting and sharing as well as to patient privacy concerns, published clinical data often contains only minimal patient information required to reproduce the results of a study [43].

Molecular and cellular/tissue level data are often distributed with minimal clinical information encoded as labels that can be used for supervised analysis (i.e., analysis that infers a function from labeled data). For example, functions that classify patients using biomarkers are often the end-product of supervised analysis of labeled biomedical data.

Common data labels include disease presence [44], disease subtype or grade [45], therapeutic regimen [46], and therapeutic outcome [47]. Thus, integration of patient level data with molecular or cellular/tissue data has long been an important part of biomedical informatics. Clinical factors that are independent of -omic and imaging data (as described in Sections II-A and II-B) are also important in disease prognosis. For example, Gleason score, tumor volume, and surgical margins are some factors that can predict survival after prostatectomy [48]; tumor stage, patient age, histologic category, and tumor grade are statistically significant factors in the International Neuroblastoma Risk Group (INRG) classification system [49]. Although clinical factors alone are predominantly used in practice, they are often limited in accuracy. For example, the Framingham Risk Score estimates cardiovascular disease risk using clinical factors, but is not applicable to nearly 15% to 20% of cardiovascular disease patients because these patients exhibit none of the traditional risk factors [50]. Thus, integration of clinical factors with other biomedical data types is necessary to improve the accuracy of diagnostic and prognostic prediction. The case study in Section III-E describes the clinical data available in the Cancer Genome Atlas data repository, as well as an example of how it can be used for integrative analysis with both -omic and imaging data.

D. Biomedical Knowledge Bases

In addition to molecular, cellular/tissue, and patient data, biomedical knowledge bases have become an important resource for integrative data analysis [51]. Vast quantities of biomedical knowledge are published in literature. Biomedical knowledge may include evidence that supports or refutes biological or clinical relationships. For example, some of these relationships could be: 1) biomarker to biomarker; 2) biomarker to biological function; or 3) biomarker to clinical outcome. However, such information is usually unstructured and difficult to mine algorithmically. Fortunately, several endeavors to organize biological knowledge have produced useful knowledge bases. For example, the GeneSigDB is a database of gene signatures that have been extracted and manually curated from published literature [52]. The Gene Expression Atlas is another curated database of biomarkers that can be queried by biological condition [53]. The Cancer Gene Census is an ongoing effort to catalogue oncogenic mutations [54]. The Cancer Gene Index (CGI) is also manually curated and contains thousands of genes linked to cancer subtypes defined in the NCI Thesaurus of controlled terminology [55], [56]. The case study in Section III-C illustrates knowledge and molecular level data integration using the CGI.

Gene ontologies can also be used as sources of biological knowledge. Perhaps the best known ontology is the Gene Ontology, which is a standard vocabulary that classifies gene products into three major classes: molecular function, biological process, and cellular component. This vocabulary is by no means exhaustive; however, it represents significant progress towards a common standard for biological annotation. The Gene Ontology Project [57] is a collaborative effort to achieve a consistent annotation for gene products across various databases. By locating selected biomarkers in the biological context of an ontological tree, evidence can be gathered to support or refute the significance of potential biomarkers. Numerous software applications are available for aiding researchers in interpreting lists of candidate genes identified from gene expression studies. For example,

AmiGO [58] allows users to store, browse and visualize ontological terms and related gene product annotation data. Likewise, Gene Ontology (GO) Explorer [59] and QuickGO [60], among many others are web-based Gene Ontology applications. In addition to the Gene Ontology, the Protein Ontology [61] is intended to facilitate the linking of genetic and proteomic function. The NCI Thesaurus and the Human Disease Ontologies catalogue thousands of cancer subtypes and other diseases. For a more comprehensive list of biomedical ontologies, refer to The Open Biological and Biomedical Ontologies website (<http://www.obofoundry.org/>) [62].

III CASE STUDIES IN BIOMEDICAL DATA INTEGRATION

Advances in biomedical data acquisition technology have enabled researchers to consider biomedical data in its entirety. As a result, there is a need to provide solutions to handle, interpret, and pursue as many aspects of the biomedical data acquired as possible. This section includes several case studies that illustrate some scenarios of integrative biomedical data analysis. Sections III-A and III-B are case studies of genomic data meta-analysis and histopathological image data normalization, respectively. Effective integration of these data types is necessary because of suboptimal data quality. Section III-C presents an example of knowledge-driven bioinformatics, an alternative method that can enhance biomedical data analysis. Sections III-D and III-E describe the HPA and TCGA, multiscale data repositories that integrate molecular, cellular/tissue, and patient data.

A. Genomic Data Meta-Analysis

Meta-analysis of gene expression data has emerged because of the increased availability of datasets coupled with the small sample size of each dataset compared to the number of features [63]. These data properties make it difficult to achieve a desired sensitivity when detecting differentially expressed genes [64]. Methods for combining multiple datasets by pooling information can increase the statistical power of biomarker identification [65]. However, combining gene expression datasets can be difficult due to technical variability arising from differences in data platform or from differences in normalization and preprocessing methods [66]. The review by Reimers details steps in microarray data analysis that include normalization and preprocessing to help researchers make informed analytical decisions. Recommended algorithms typically depend on the microarray technology [67]. The analysis by Lim *et al.* focuses on normalization methods for Affymetrix microarrays and suggests that MAS5 works well for their application of reverse engineering gene networks; however, other normalization methods (e.g., RMA and GCRMA) may be better choices for applications such as detection of differentially expressed genes [68]. Even when the same technology, normalization, and preprocessing software are used, the laboratory effect can introduce significant variance [69], [70]. This case study presents an example of meta-analysis using five renal cancer gene expression datasets, describes the challenges involved with both inter- and intraplatform data combination, and illustrates some solutions to these challenges.

Although the renal tumor gene expression datasets in Table I share a similar clinical focus, they were acquired with different microarray technology [71], [72], [73], [74], [75]. Because

different microarray models use different probe sets, interplatform probe matching is necessary for combined-data analysis. The Affymetrix microarray datasets (i.e., Schuetz, Jones, Kort, and Yusenko) can be reduced to a common set of 8793 probe sets. However, the Higgins dataset contains 22689 cDNA probes. Matching probes between heterogeneous microarray platforms may be achieved by either annotation-based (i.e., matching gene names or symbols) or sequence-based (i.e., matching probe sequences) methods. Sequence-based probe matching tends to result in better consistency [76]. Thus, the Batch Entrez tool (<http://www.ncbi.nlm.nih.gov/sites/batchentrez>) was used to download EST sequences for all cDNA probes in the Higgins dataset. These sequences were then aligned to the exemplar and consensus sequences (obtained from <http://www.affymetrix.com/>) of the 8793 Affymetrix probes using the Burrows-Wheeler Aligner [77]. Identification of all unique one-to-one matches from the sequence alignment resulted in 946 common probes.

After probe matching, differences in technology can still result in differences in dynamic range of gene expression. The removal of technical biases in biomedical data, often called batch effects, is possible and numerous studies have proposed novel strategies [78], [79]. The quantile normalization method, for example, can force the distribution of the Higgins cDNA dataset to be similar to that of the Affymetrix datasets [80]. Biological and laboratory variance can also cause batch effects even when the data platform and clinical focus are similar.

Consider the case of the Schuetz and Jones renal tumor gene expression datasets, each acquired at different laboratories [71], [72]. Despite the fact that both datasets measure similar biological samples (i.e., similar renal tumor subtypes and similar patient groups) using similar Affymetrix technology, the distribution of gene expression in each microarray sample is dataset dependent [Fig. 2(a)]. Quantile normalization removes the batch effect in gene expression distribution (Fig. 2(a), red line). However, in some cases, batch effect correction does not completely solve the problem or can have undesirable consequences [81]. As seen in Fig. 2(b), after quantile normalization, some differences in gene expression between the two datasets remain dataset dependent. The primary clusters of hierarchical clustering correlate with the datasets rather than with the renal tumor subtypes.

Gene expression meta-analysis methods that combine dataset statistics rather than dataset samples may be more robust to batch effects. For example, some methods model the variance of gene expression among several datasets in order to identify unwanted variance (e.g., laboratory or technology variance) and improve differential gene detection [82], [83]. Other studies compute statistics such as fold changes or gene ranks for each individual dataset and combine these statistics [65], [84].

Another solution to the problem of heterogeneous biomedical data attempts to avoid batch effects altogether by mandating standard experimental protocols for data acquisition. Special interest groups in the biomedical engineering community are actively developing experimental and documentation standards to facilitate unambiguous communication of genomic experiments [85]. For instance, the Microarray and Gene Expression Data Society (MGED) has established several standards to facilitate sharing of high-throughput molecular gene expression data. One of their standards, the Minimum Information About a Microarray

Experiment (MIAME), provides all information necessary to enable unambiguous interpretation and reproduction of experimental microarray results [85]. This standard is used by several large data repositories, including ArrayExpress [86] and the Gene Expression Omnibus (GEO) [87]. However, as described in [85], MIAME is only a set of guidelines and does not provide a data format. Implementation of the guidelines into a data format with supporting software tools proved to be challenging. Thus, different interpretations of the guidelines were implemented at various data repositories, resulting in difficulties in reproducing research results. Eventually, data repositories began to adopt a simpler (and more easily implemented) MIAME compliant MAGE-TAB (microarray gene expression tabular) format [88].

Meta-analysis and data combination are applicable not only to molecular level data, but also to cellular and tissue level data. Regardless of the biomedical data type, such integrative analyses are subject to similar obstacles, e.g., batch effects. The following case study illustrates a method for batch effect normalization of histopathological image data.

B. Histopathological Image Data Normalization

Computer-aided histopathological image classification has become an important area of biomedical research because of the potential improvement that it brings to traditional diagnostic pathology. For applications such as cancer subtyping or grading, computer-aided prediction systems rely on image features such as nuclear shape or topology. However, much like genomic data, histopathological image data can suffer from technological batch effects that are problematic in clinical prediction applications, especially when data are expected to come from different laboratories and to be produced using different protocols. Thus, normalization of histopathological images is important for integrative analysis at the cellular and tissue level.

Scale and color batch effects are common in histopathological images. Scale batch effects occur because of variations in microscope characteristics (e.g., numerical aperture, magnification, or digitizing device) that can affect image features such as nuclear size, topology, and texture. Color batch effects occur because of variations in specimen preparation, staining, and imaging that can affect color-based image features in the RGB [89], [90], [91] or HSV color spaces [92], as well as color-based segmentation methods. Color-based segmentation is essential for extracting object-based properties of the cellular structures in histopathological samples (e.g., nuclei, cytoplasm, leukocytes, red blood cells, and glands) [93], [94], [95], [96]. These cellular structures enable clinicians to identify morphological markers of disease and to proceed with therapy accordingly. Thus, the accuracy of image segmentation methods affects image features and, subsequently, the robustness of computer-aided diagnostic systems. This case study focuses on color normalization techniques for histopathological image data.

Several color normalization techniques have been proposed, including histogram or quantile normalization in which the distributions of the three color channels are normalized separately [96], [97], [98]. The case study illustrated in Fig. 3 compares two color normalization approaches: quantile normalization of all image pixels and quantile normalization of only unique colors [96]. In Fig. 3(a), two renal tumor histopathological

images differ significantly in color distribution. Pixels in the reference image tend to have much higher blue-channel components compared to the test image, resulting in an image that appears brighter. Points in the scatter plots are colored to represent different cellular structures: blue represents nuclei, pink represents cytoplasm, black represents glands or no-stain, and red represents red-blood cells. Without color normalization, segmentation of the test image using the reference image (as described in [96]) results in low segmentation accuracy of 25%. In the color space, all pixels in the test image are close to the nuclei of the reference image and are mostly incorrectly classified as nuclei. Normalization of the individual red, green, and blue channels of all pixels in the test image to that of the reference image improves segmentation accuracy to 68% [Fig. 3(b)]. However, the morphologies of the reference and test images are very different, resulting in different frequencies of structure-specific pixels. For example, the reference image has more white space and less cytoplasm compared to the test image because of differences in renal cancer subtype morphology. Thus, normalization using all pixels forces the normalized image to have similar frequencies of structure-specific pixels, resulting in altered morphology and lower segmentation accuracy. Normalization of the individual red, green, and blue channels of only the unique pixels (in terms of RGB color) improves segmentation accuracy to 79% [Fig. 3(c)]. Normalization of only the unique pixels is independent of the frequencies of structure-specific pixels and preserves the morphology of the test image.

The ability to combine multiple histopathological datasets regardless of technological batch effects is necessary for robust computer-aided diagnostic systems. Effective algorithms for normalization of both color (illustrated in this case study) and scale can improve the accuracy of such systems by increasing the number of samples available for predictive modeling. An alternative method to improve the effectiveness of integrative analysis of biomedical data involves the use of domain or biological knowledge bases to guide algorithm selection as described in the following case study.

C. Knowledge-Driven Bioinformatics

The case studies presented in Sections III-A and III-B showed that molecular and cellular/tissue level data often suffer from small sample size compared to the relatively large information content of each sample (e.g., tens of thousands of genes or millions of image pixels). Such data sparseness in some dimensions coupled with data abundance in other dimensions gives rise to a common biomedical data problem known as the “curse of dimensionality” [51]. This case study first illustrates the curse of dimensionality, and then describes how knowledge-driven methods (as alternative or complimentary methods to meta-analysis or data combination) can potentially improve results under ill-posed data conditions.

The curse of dimensionality describes an exponential increase in mathematical problem space that occurs when the problem dimension increases. The increase in problem space reduces the overall density of data samples. In data driven decision-making applications that govern biomedical data analysis, low data density can result in lower statistical confidence. Consider a 1-D data situation in which all data points are evenly spaced on an interval of unit length. Hypothetically, this could represent the normalized expression of a single gene

measured from multiple patients. If the density of samples is required to meet a specific criterion (e.g., adjacent samples should be spaced at no more than 0.2 units), five samples are necessary at one dimension. Increasing the data dimension to two—which could represent the joint expression of two genes—increases the number of samples required to satisfy the criterion to 25 samples. At three dimensions, this number increases to 125 samples (Fig. 4). Unfortunately, practical data analysis scenarios for biomedical data analysis are extreme cases of the curse of dimensionality. It is impractical and statistically unsound to analyze bio-molecular or bio-imaging data at their natural dimension. Thus, feature extraction and selection methods are often used to alleviate the curse of dimensionality problem. However, feature extraction and selection are difficult because of data dimension (i.e., exponential number of feature combinations) and variance resulting from small sample size and numerous algorithm choices.

Knowledge-driven bioinformatics methods are promising solutions to the computational and curse-of-dimensionality problems [51]. In biomarker identification applications, it is often difficult for researchers to choose an appropriate algorithm due to the sheer number of available algorithms. Previous studies have developed metrics to compare and choose algorithms using domain knowledge. For example, Mukherjee *et al.* developed a theoretical framework to compare microarray gene ranking metrics in the presence of control genes [99]. Chen similarly compared an independent component analysis (ICA) method to their knowledge-guided method and found that the knowledge-guided method improved biomarker detection efficiency [100]. Phan extended the framework introduced by Mukherjee and used laboratory and clinically validated biomarkers as knowledge to improve biomarker detection efficiency [101].

As an example of knowledge-driven algorithm selection, consider the Schuetz renal tumor gene expression dataset introduced in Section III-A. Several feature ranking methods (e.g., rank products [84], fold-change, SAM [102], mRMR [103], T-test, and rank sum) can be used to identify biomarkers that can classify the microarray samples into one of four renal tumor subtypes (i.e., clear cell, chromophobe, papillary, and oncocytoma). However, without some guidance of domain knowledge (i.e., literature- or laboratory-verified biomarkers for renal tumor subtypes) it is unclear which of the feature ranking methods is most biologically relevant. Several biomarker knowledge bases are available that can be used as a “ground truth” for assessing the biological relevance of feature selection algorithms (some knowledge bases are described in Section II-D). Using the Cancer Gene Index (CGI) as a knowledge base, several probes in the Schuetz microarray dataset can be linked to CGI genes related to renal tumor subtypes (Table II). These probes can be used to compute the biological relevance of the seven feature selection methods in Table III. The biological relevance, computed as in [101], is the probability that microarray probes linked to the knowledge base are ranked more favorably than other probes. There is a clear difference in biological relevance between the rank products (0.69) and rank sum (0.54) methods. Thus, according to [101], the rank products method should be able to better identify novel and biologically relevant biomarkers from this renal tumor subtyping dataset.

D. Human Protein Atlas

Section II-B introduced immunohistochemistry (IHC) and immunofluorescence (IF) imaging as cellular and tissue level biomedical data types that inherently integrate molecular level data. The case study in this section describes the content and potential applications of the Human Protein Atlas (HPA, <http://www.proteinatlas.org/>), a large public data repository that integrates molecular level data (in the form of protein expression) with cellular and tissue level data for a variety of diseased and normal human tissues [104]. The HPA includes extensive profiles for each protein across 48 normal tissues, 20 cancer types, and 59 cell lines using digitized images and aggregated information about protein expression levels. The immunoreactivity of each protein is evaluated and annotated by certificated pathologists. Currently, the HPA contains more than one million images corresponding to more than 10000 in-house and commercially available human protein antibodies. Thus, the HPA can be used as a resource for validating molecular or cellular/tissue biomarkers for various diseases.

Each protein-coding gene in the HPA is labeled using the Ensembl annotation system and linked to gene names, descriptions, chromosomal locations, and external databases (e.g., NCBI Entrez Gene and Uniprot). Moreover, antibody identifiers and results for validation of protein expression are also included. Protein expression validation methods include IHC staining, IF, western blots, and protein arrays. Fig. 5 is an example of some of the protein expression data in the HPA for BRCA1, a gene implicated in breast cancer. The summary page for BRCA1 lists three antibodies for its corresponding protein: HPA34966, CAB001946, and CAB018369. IF staining using the HPA034966 antibody indicates that the protein product of BRCA1 is localized in the plasma membrane, cytoplasm, and nucleus of the U-2 OS human cell line [Fig. 5(a)]. Cytoplasmic and nuclear localization are corroborated with IHC staining of normal human breast tissue [Fig. 5(b)]. IHC staining results for many other normal human and cancer tissues are also included. Fig. 5(c) lists 20 cancer types with the level of BRCA1 antibody staining, with red indicating fraction of IHC experiments showing strong staining and white indicating fraction of experiments showing negative staining.

Applications of the HPA include discovery and validation of disease biomarkers. Section II-B discussed the importance and difficulty of monitoring protein activity, and Section II-C highlighted the importance of spatial information in protein expression. The HPA provides a means to link gene expression profiles to protein expression and localization profiles [105]. For example, researchers can use the HPA to find evidence supporting candidate genomic biomarkers by verifying the expression of the gene's protein product. However, inability to verify the expression of a protein product does not eliminate the corresponding gene as a potential biomarker. Moreover, the HPA may be used as a knowledge base in a similar manner to that described in Section III-C. Protein-coding genes with strong evidence of differential protein expression in specific tissue types may be used as knowledge to guide algorithm selection for biomarker discovery.

E. The Cancer Genome Atlas

TCGA is a molecular, cellular/tissue, and patient data repository that aims to provide a comprehensive resource for integrative cancer research. It contains up to 500 patient samples

for each of the 20 included cancers. Each patient sample is assayed with whole-slide tissue imaging, gene expression microarrays, and several other genomic and proteomic quantification technologies. Although data collection is ongoing, publications using TCGA data have conducted integrative analysis of glioblastoma [29], [38], [106], ovarian cancer [107], [108], and lung adenocarcinoma [109] data. This case study summarizes the methods and findings of one such publication in order to illustrate the data and analytical possibilities provided by TCGA.

The study by Cooper *et al.* illustrates the progress and possibilities of integrating multiscale biomedical data from morphological (i.e., cellular/tissue imaging) to clinical to genomic data [29]. In this study, features of glioblastoma extracted from whole-slide images are analyzed to determine associations between representative models of tumor morphology (i.e., visual tumor signatures), clinical indicators (e.g., patient outcome, response to therapy), and underlying molecular events (i.e., genetic alterations). In particular, the authors tested this pipeline using: 1) morphological data, in terms of 462 whole-slide histopathological images of 162 glioblastomas from TCGA that comprises 200 million cells as well as pathology classifications [Fig. 6(a)]; 2) clinical data, in terms of patient outcome (i.e., survival taken as “days to death”, and response to intensive therapy detailed as three or more cycles of chemotherapy or concurrent radiation and chemotherapy, [Fig. 6(b)]; 3) genomic data, transcriptional class labels, methylation phenotype, and genetic alterations of recognized genes, in terms of microarray gene expression using Affymetrix, Agilent, and Illumina platforms [Fig. 6(c)]; and 4) molecular pathway associations using Gene Ontology. The results were validated with independent glioblastoma data from the Henry Ford health system. Thus, the authors implemented and tested an integrative morphological/genomic analysis pipeline that consists of four modules: 1) morphology; 2) clustering; 3) correlative; and 4) genome-wide analysis.

From this integrative morphological/genomic analysis using TCGA glioblastoma data, the authors identified three clusters from tumor morphology, which could also be associated with specific genes involved in biological functions such as cell cycle, chromatin modification, and protein biosynthesis. Furthermore, the prognosis of patients identified from these morphological clusters could also be differentiated statistically. The scope of the integrative morphological/genomic analysis pipeline also enabled associations between morphological clusters, cellular pathways, gene expression, differential methylation, copy number, and expression mapping to be identified. However, validation with independent data confirmed only two of the three clusters even though these also exhibited similar survival trends with the original data. Furthermore, the morphological clusters were not strongly associated with traditional neuropathology classification or genetic alterations of recognized genes. While these results illustrate the progress and possibilities of multiscale biomedical data integration, they also reveal several interesting pitfalls of an integrative data analysis framework.

First, while morphological clusters can also be associated with differences in prognosis, not all of these clusters can be identified in different data as seen from the authors’ validation with independent data. This is an interesting scientific question that may be a result of limited training data. At the same time, what is considerably more challenging is that the

clinical relevance or functionality of these differences (i.e., from the perspective of treating the patient) remains unclear. In other words, how can this information be used to guide and inform a patient's treatment?

Second, in terms of data integration, there is first a reduction phase (to minimize the number of "essential" features), then an expansion phase (to recapture empirical conditions). Current results based on representative signatures of patient-specific tumor morphology may not capture heterogeneous cell populations in the tumor microenvironment. As a result, there is a need to expand the number of features for consideration (e.g., expansion to include molecular gene expression features). In addition, estimating representative signature models from data feature extraction and integration is an iterative process that requires many computing cycles to achieve reasonable convergence (i.e., high computational costs are involved). Consequently, investigation of meta-analysis frameworks to efficiently manage such integrative models is important.

Significant, but unavoidable, challenges have emerged as a result of the large scale of the TCGA endeavor. Ongoing research to overcome these challenges has led to improvements both in bioinformatics infrastructure as well as in overall understanding of cancer. First, the enormous size of the data, especially the image and sequence data, has spurred the development of new computational infrastructure capable of storing, analyzing, and sharing such large data. For example, high-resolution histopathological images may contain hundreds of millions of pixels and next-generation sequencing samples may contain tens of millions of reads. Histopathological images must often be cropped into smaller sections, or tiles, in order to achieve efficient processing and visualization [110]. Similar methods for visualization of large genomic datasets have also been developed [111]. Second, data quality may not be perfect. For example, tissue image morphology may be unclear in large portions of TCGA images because of the image acquisition protocol (i.e., frozen versus paraffin-embedded tissue sections). Thus, low-quality portions of images (i.e., portions with unclear morphology) must be identified and removed prior to extraction of features for diagnostic purposes. Moreover, development of image data mining methods that consider a large variety of features may identify useful features in situations where traditional morphological features are not clearly distinguishable [90]. Finally, while there is undoubted progress in the community to integrate multiscale biomedical data, the challenge remains to harness such research outcomes to improve clinical outcomes. Moreover, it is not clear which pieces of TCGA data will lead to the most clinically useful answers. TCGA provides a valuable resource of multiscale patient data for exploring integrative analysis methods with potential clinical impact for cancer diagnosis, prognosis, and therapy. The public availability of TCGA data has stimulated a wide variety of integrative data analysis methods, resulting in interesting cancer biomarker discoveries [112], [113] as well as new bioinformatics software tools [111], [114].

IV CONCLUSION

This review has presented the benefits and challenges of multiscale biomedical data integration in the context of personalized, predictive, and preventive medicine. There is a need for improved efficacy in terms of: 1) predictive disease diagnosis; 2) personalized

therapeutic decisions; and 3) preventive healthcare. Biomedical informatics is essential for realizing these medical applications. Because of the wide variety and heterogeneity of biomedical data sources, multiscale data integration is an important component of biomedical informatics. Research opportunities and challenges remain for multiscale data integration in terms of data heterogeneity, data standards, and computational complexity.

Biomedical data complexity and heterogeneity are the primary challenges of multiscale data integration. As discussed in Section II, there are many biomedical data acquisition technologies covering several data scales (i.e., molecular, cellular/tissue, patient, and knowledge-based data). While complexity naturally arises from multiple scales of data acquisition, heterogeneity arises due to the lack of standards in data acquisition technology as well as to the variability in experimental protocols.

The challenge of data heterogeneity is pervasive in biomedical data analysis. It is often seen in molecular level data (although not limited to molecular level data) because of numerous competing acquisition technologies that are not necessarily compatible. Even data acquired with similar technology may not be directly comparable due to inconsistencies in experimental protocol. In the bioinformatics literature, methods for dealing with data heterogeneity are equally pervasive. The case studies in Sections III-A and III-B presented some methods for handling data heterogeneity in both the molecular and cellular/tissue levels. Other solutions to data heterogeneity aim to properly standardize minimum information requirements to achieve data reproducibility (see Section III-A). The lack of common standards for biomedical data integration impedes researchers, software developers, and users from sharing solutions and collaborating effectively. Thus, research opportunities exist for development of standards and methods for biomedical data acquisition.

Biomedical data integration is inherently computationally complex due to the sparseness of data in some dimensions and the abundance of data in other dimensions. The case study in Section III-C described this phenomenon in detail and presented a knowledge-based method to improve the analysis of sparse biomedical data. Despite the computational complexity, biomedical data integration has become increasingly prevalent. Large multiscale data repositories such as the HPA (Section III-D) and TCGA (Section III-E) have emerged as a result of the increasing emphasis on integrative analysis.

Finally, this review has focused on case studies involving molecular, cellular/tissue, and patient level data as examples of some of the possible multiscale data integration scenarios. However, other levels of data exist that are also important for personalized, predictive, and preventive medicine. These include organ level data (e.g., medical imaging) and physiological level data (e.g., blood pressure, electrocardiography, body temperature, etc.). Inclusion of these data levels could improve the overall benefit of multiscale data integration for biomedical informatics.

References

1. Hunter PJ, Viceconti M. The VPH-physiome project: Standards and tools for multiscale modeling in clinical applications. *IEEE Rev Biomed Eng.* 2009; 2:40–53.

2. Straver ME, Glas AM, Hannemann J, Wesseling J, van de Vijver MJ, Rutgers EJT, Peeters MJTFDV, van Tinteren H, van 't Veer LJ, Rodenhuis S. The 70-gene signature as a response predictor for neoadjuvant chemotherapy in breast cancer. *Breast Cancer Res Treatment*. 2010; 119:551–558.
3. Sorensen KD, Orntoft TF. Discovery of prostate cancer biomarkers by microarray gene expression profiling. *Expert Rev Mol Diagn*. 2010; 10:49–64.
4. Bratslavsky G, Sanford T, Srinivasan R, Aprelikova O, Liu J, Quezado M, Merino M, Linehan WM. Differential genetic expression in large versus small clear cell renal cell carcinoma: Results from microarray analysis. *J Cancer*. 2011; 2:271–279. [PubMed: 21611108]
5. Majewski IJ, Bernards R. Taming the dragon: Genomic biomarkers to individualize the treatment of cancer. *Nat Med*. 2011; 17:304–312. [PubMed: 21386834]
6. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM. NCBI GEO: Archive for functional genomics data sets—10 years on. *Nucleic Acids Res*. 2011; 39:D1005–D1010.
7. Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Research*. 2011; 39:D19–D21. [PubMed: 21062823]
8. Quackenbush J. Microarray analysis and tumor classification. *N Engl J Med*. 2006; 354:2463–2472. [PubMed: 16760446]
9. Heller MJ. DNA microarray technology: Devices, systems, and applications. *Annu Rev Biomed Eng*. 2002; 4:129–153. [PubMed: 12117754]
10. Schulze A, Downward J. Navigating gene expression using microarrays—A technology review. *Nature Cell Biol*. 2001; 3:E190–E195.
11. Metzker ML. Sequencing technologies—The next generation. *Nat Rev Genet*. 2009; 11:31–46. [PubMed: 19997069]
12. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009; 10:57–63. [PubMed: 19015660]
13. Gorlov IP, Gallick GE, Gorlova OY, Amos C, Logothetis CJ. GWAS meets microarray: Are the results of genome-wide association studies and gene-expression profiling consistent? Prostate cancer as an example. *PLoS One*. 2009; 4:e6511–e6511. [PubMed: 19652704]
14. Luo L, Boerwinkle E, Xiong M. Association studies for next-generation sequencing. *Genome Res*. 2011; 21:1099–1108. [PubMed: 21521787]
15. VanDe Wiel MA, Picard F, Van Wieringen WN, Ylstra B. Preprocessing and downstream analysis of microarray DNA copy number profiles. *Briefings in Bioinformatics*. 2011; 12:10–21.
16. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics*. 2009; 41:1061–1067. [PubMed: 19718026]
17. Sandoval J, Heyn HA, Moran S, Serra-Musach J, Pujana MA, Bibikova M, Esteller M. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics: Official J DNA Methylation Soc*. 2011; 6
18. Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, Egli D, Maherali N, Park IH, Yu J. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature Biotechnol*. 2009; 27:353–360.
19. Sato F, Tsuchiya S, Terasawa K, Tsujimoto G. Intra-platform repeatability and inter-platform comparability of microRNA microarray technology. *PLoS One*. 2009; 4:e5540–e5540. [PubMed: 19436744]
20. Buermans HPJ, Ariyurek Y, van Ommen G, den Dunnen JT, Hoen PA. New methods for next generation sequencing based expression profiling. *BMC Genomics*. 2010; 11:716–716.
21. Gstaiger M, Aebersold R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat Rev Genet*. 2009; 10:617–627. [PubMed: 19687803]
22. Nilsson T, Mann M, Aebersold R, Yates JR, Bairoch A, Bergeron JJM. Mass spectrometry in high-throughput proteomics: Ready for the big time. *Nature Methods*. 2010; 7:681–681. [PubMed: 20805795]
23. Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: Approaches, advances, and applications. *Annu Rev Biomed Eng*. 2009; 11:49–79. [PubMed: 19400705]

24. Ramachandran N, Raphael JV, Hainsworth E, Demirkan G, Fuentes MG, Rolfs A, Hu Y, LaBaer J. Next-generation high-density self-assembling functionalprotein arrays. *Nature Methods*. 2008; 5:535–538. [PubMed: 18469824]
25. Hudson ME, Pozdnyakova I, Haines K, Mor G, Snyder M. Identification of differentially expressedproteins in ovarian cancer using high-density protein microarrays. *Proc Nat Acad Sci*. 2007; 104:17494–17494. [PubMed: 17954908]
26. Novara G, Martignoni G, Artibani W, Ficarra V. Grading systems in renal cellcarcinoma. *J Urol*. 2007; 177:430–436. [PubMed: 17222604]
27. Humphrey PA. Gleason grading and prognosticfactors in carcinoma of the prostate. *ModernPathol*. 2004; 17:292–306.
28. Malpica A, Deavers MT, Lu K, Bodurka DC, Atkinson EN, Gershenson DM, Silva EG. Grading ovarian serous carcinoma using atwo-tier system. *Amer J Surgical Pathol*. 2004; 28:496–496.
29. Cooper LAD, Kong J, Gutman DA, Wang F, Gao J, Appin C, Cholleti S, Pan T, Sharma A, Scarpace L. Integrated morphologic analysis for theidentification and characterization of disease subtypes. *J Amer Med Informatics Assoc*. 2012; 19:317–323.
30. Resch-Genger U, Grabolle M, Cavaliere-Jaricot S, Nitschke R, Nann T. Quantum dots versus organic dyes as fluorescentlabels. *Nat Methods*. 2008; 5:763–775. [PubMed: 18756197]
31. Gao X, Cui Y, Levenson RM, Chung LWK, Nie S. Vivo cancer targeting and imaging with semiconductor quantumdots. *Nature Biotechnol*. 2004; 22:969–976. [PubMed: 15258594]
32. Tholouli E, Sweeney E, Barrow E, Clay V, Hoyland JA, Byers RJ. Quantum dots light up pathology. *J Pathol*. 2008; 216:275–285. [PubMed: 18814189]
33. Pinaud F, Clarke S, Sittner A, Dahan M. Probing cellular events, one quantum dotat a time. *Nature Methods*. 2010; 7:275–285. [PubMed: 20354518]
34. Chin PTK, Buckle T, de Miguel AA, Meskers SCJ, Janssen RAJ, van Leeuwen FWB. Dual-emissive quantum dotsfor multispectral intraoperative fluorescence imaging. *Biomaterials*. 2010; 31:6823–6832. [PubMed: 20619786]
35. Schwamborn K, Caprioli RM. Molecular imaging by mass spectrometry—Lookingbeyond classical histology. *Nat Rev Cancer*. 2010; 10:639–646. [PubMed: 20720571]
36. Chughtai K, Heeren RMA. Mass spectrometric imagingfor biomedical tissue analysis. *Chem Rev*. 2010; 110:3237–3237. [PubMed: 20423155]
37. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S. Towards a knowledge-based human protein atlas. *Nat Biotechnol*. 2010; 28:1248–1250. [PubMed: 21139605]
38. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, Olson JJ, Mikkelsen T, Lehman N, Aldape K. Comprehensive genomic characterizationdefines human glioblastoma genes and core pathways. *Nature*. 2008; 455:1061–1068. [PubMed: 18772890]
39. Bell D, Berchuck A, Birrer M, Chien J, Cramer D, Dao F, Dhir R, Disaia P, Gabra H, Glenn P. Integrated genomic analysesof ovarian carcinoma. *Nature*. 2011; 474:609–615. [PubMed: 21720365]
40. Rasmussen LV, Peissig PL, McCarty CA, Starren J. Development of an optical characterrecognition pipeline for handwritten form fields from an electronic healthrecord. *J Amer Med Informatics Assoc*. 2011
41. Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, Shields A, Rosenbaum S, Blumenthal D. Use of electronic health recordsin US hospitals. *N Engl J Medicine*. 2009; 360:1628–1638.
42. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, Shvo AS. HL7 clinical document architecture, release 2. *J Amer Med Informatics Assoc*. 2006; 13:30–39.
43. Hrynaskiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical datafor publication: Guidance for journal editors, authors, and peer reviewers. *BMJ (CR)-Print*. 2010; 340:181–181.
44. Cao DL, Ye DW, Zhang HL, Zhu Y, Wang YX, Yao XD. A multiplex model of combining gene-based, protein-based, and metabolite-based with positive and negative markers in urine for the earlydiagnosis of prostate cancer. *The Prostate*. 2011; 71:700–710. [PubMed: 20957673]
45. Chow WH, Dong LM, Devesa SS. Epidemiology and risk factorsfor kidney cancer. *Nature Rev Urol*. 2010; 7:245–257. [PubMed: 20448658]

46. Rajkumar T, Vijayalakshmi N, Sabitha K, Shirley S, Selvaluxmy G, Bose MV, Nambaru L. A 7 gene expression score predicts for radiation response in cancer cervix. *BMCCancer*. 2009; 9:365–365.
47. Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, Lopez-Doriga A, Santos C, Marijnen C, Westerga J. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol*. 2011; 29:17–24. [PubMed: 21098318]
48. Buhmeida A, Pyrhonen S, Laato M, Collan Y. Prognostic factors in prostate cancer. *Diagn Pathol*. 2006; 1:124–124.
49. Cohn SL, Pearson ADJ, London WB, Monclair T, Ambros PF, Brodeur GM, Faldut A, Hero B, Iehara T, Machin D. The International Neuroblastoma Risk Group (INRG) classification system: An INRG task force report. *J Clin Oncol*. 2009; 27:289–297. [PubMed: 19047291]
50. Thanassoulis G, Vasan RS. Genetic cardiovascular risk prediction will we get there. *Circulation*. 2010; 122:2323–2334. [PubMed: 21147729]
51. Phan JH, Moffitt RA, Stokes TH, Liu J, Young AN, Nie S, Wang MD. Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment. *Trends Biotechnol*. 2009; 27:350–358.
52. Culhane AC, Schröder MS, Sultana R, Picard SC, Martinelli EN, Kelly C, Haibe-Kains B, Kapushesky M, Pierre AAS, Flahive W. GeneSigDB: A manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Res*. 2012; 40:D1060–D1066.
53. Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, Holloway E, Klebanov A, Kryvych N, Kurbatova N. Gene Expression Atlas update—A value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res*. 2011
54. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. *Nature Rev Cancer*. 2004; 4:177–183. [PubMed: 14993899]
55. Schueller C, Fritz A, Torres S, Wenger K, Albermann K, Komatsoulis G, Covitz P, Wright L, Hartel F. Towards a comprehensive catalog of gene-disease and gene-drug relationships in cancer. *Proc Int Conf Intelligent Systems for Molecular Biology*. 2005
56. Sioutos N, Coronado S, Haber M, Hartel F, Shaiu W, Wright L. NCI thesaurus: A semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform*. 2007; 40:30–43. [PubMed: 16697710]
57. Harris M, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 2004; 32:D258–D261.
58. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. AmiGO: Online access to ontology and annotation data. *Bioinformatics*. 2009; 25:288–289. [PubMed: 19033274]
59. Carvalho PC, Fischer JSG, Chen EI, Domont GB, Carvalho MGC, Degraive WM, Yates JR, Barbosa VC. GO explorer: A gene-ontology tool to aid in the interpretation of shotgun proteomics data. *Proteome Sci*. 2009; 7:6–6. [PubMed: 19239707]
60. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics*. 2009; 25:3045–3045. [PubMed: 19744993]
61. Natale D, Arighi C, Barker W, Blake J, Chang T, Hu Z, Liu H, Smith B, Wu C. Framework for a protein ontology. *BMC Bioinform*. 2007; 8:S1–S1.
62. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnol*. 2007; 25:1251–1255. [PubMed: 17989687]
63. Gadaleta E, Lemoine NR, Chelala C. Online resources of cancer data: Barriers, benefits and lessons. *Brief Bioinform*. 2011; 12:52–63.
64. Lin WJ, Hsueh HM, Chen JJ. Power and sample size estimation in microarray studies. *BMC Bioinform*. 2010; 11
65. Campaign A, Yang Y. Comparison study of microarray meta-analysis methods. *BMC Bioinform*. 2010; 11:408–408.
66. Freyhult E, Landfors M, Önskog J, Hvidsten T, Rydén P. Challenges in microarray class discovery: A comprehensive examination of normalization, gene selection and clustering. *BMC Bioinform*. 2010; 11:503–503.

67. Reimers M. Making informed choices about microarray data analysis. *PLoS Computational Biol.* 2010; 6:e1000786–e1000786.
68. Lim WK, Wang K, Lefebvre C, Califano A. Comparative analysis of microarray normalization procedures: Effects on reverse engineering gene networks. *Bioinformatic.* 2007; 23:i282–i288.
69. Irizarry R, Warren D, Spencer F, Kim I, Biswal S, Frank B, Gabrielson E, Garcia J, Geoghegan J, Germino G. Multiple-laboratory comparison of microarray platforms. *Nat Methods.* 2005; 2:345–350. [PubMed: 15846361]
70. Stafford P, Brun M. Three methods for optimization of cross-laboratory and cross-platform microarray expression data. *Nucleic Acids Res.* 2007; 35:e72–e72.
71. Schuetz AN, Yin-Goen Q, Amin MB, Moreno CS, Cohen C, Hornsby CD, Yang WL, Petros JA, Issa MM, Pattaras JG, Ogan K, Marshall FF, Young AN. Molecular classification of renal tumors by gene expression profiling. *J Mol Diagn.* 2005; 7:206–218. [PubMed: 15858144]
72. Jones J, Otu H, Spentzos D, Kolia S, Inan M, Beecken WD, Fellbaum C, Gu X, Joseph M, Pantuck AJ, Jonas D, Libermann TA. Gene signatures of progression and metastasis in renal cell cancer. *Clin Cancer Res.* 2005; 11:5730–5739. [PubMed: 16115910]
73. Kort E, Farber L, Tretiakova M, Petillo D, Furge K, Yang X, Cornelius A, Teh B. The E2f3-oncomir-1 axis is activated in Wilms' tumor. *Cancer Res.* 2008; 68:4034–4038. [PubMed: 18519660]
74. Yusenko M, Kuiper R, Boethe T, Ljungberg B, van Kessel A, Kovacs G. High-resolution DNA copy number and gene expression analyses distinguish chromophobe renal cell carcinomas and renal oncocytomas. *BMC Cancer.* 2009; 9:152–152. [PubMed: 19445733]
75. Higgins J, Shinghal R, Gill H, Reese J, Terris M, Cohen R, Fero M, Pollack J, Van De Rijn M, Brooks J. Gene expression patterns in renal cell carcinoma assessed by complementary DNA microarray. *Amer J Pathol.* 2003; 162:925–932. [PubMed: 12598325]
76. Kuo WP, Liu F, Trimarchi J, Punzo C, L M, Sarang J, Whipple ME, Maysuria M, Serikawa K, Lee SY. A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nature Biotechnol.* 2006; 24:832–840. [PubMed: 16823376]
77. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatic.* 2009; 25:1754–1760.
78. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C. Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS One.* 2011; 6:e17238–e17238.
79. Johnson W, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007; 8:118–127. [PubMed: 16632515]
80. Bolstad B, Irizarry R, Astrand M, Speed T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 2003; 19:185–193. [PubMed: 12538238]
81. Lusa L, McShane L, Reid J, De Cecco L, Ambrogi F, Biganzoli E, Gariboldi M, Pierotti M. Challenges in projecting clustering results across gene expression-profiling datasets. *J Nat Cancer Inst.* 2007; 99:1715–1723. [PubMed: 18000217]
82. Park T, Yi S, Shin Y, Lee S. Combining multiple microarrays in the presence of controlling variables. *Bioinformatics.* 2006; 22:1682–1689. [PubMed: 16705015]
83. Choi J, Yu U, Kim S, Yoo O. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics.* 2003; 19:i84–i90. [PubMed: 12855442]
84. Breitling R, Herzyk P. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *J Bioinform Comput Biol.* 2005; 3:1171–1190. [PubMed: 16278953]
85. Brazma A. Minimum information about a microarray experiment (MIAME)—Successes, failures, challenges. *Scientific World J.* 2009; 9:420–423.
86. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A. ArrayExpress update—From an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.* 2009; 37:D868–D872.

87. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA. NCBI GEO: Archive for high-throughput functional genomic data. *Nucleic Acids Res.* 2009; 37:D885–D890. [PubMed: 18940857]
88. Rayner T, Rocca-Serra P, Spellman P, Causton H, Farne A, Holloway E, Irizarry R, Liu J, Maier D, Miller M. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinform.* 2006; 7:489–489.
89. Tabesh A, Teverovskiy M, Ho-Yuen P, Kumar VP, Verbel D, Kotsianti A, Saidi O. Multifeature prostate cancer diagnosis and Gleason grading of histological images. *IEEE Trans Med Imag.* 2007; 26(6):1366–1378.
90. Kothari S, Phan JH, Young AN, Wang MD. Histological image feature mining reveals emergent diagnostic properties for renal cancer. *Proc IEEE Int Conf Bioinform Biomedicine (BIBM).* 2011
91. Fuchs T, Wild P, Moch H, Buhmann J. Computational pathology analysis of tissue microarrays predicts survival of renal clear cell carcinoma patients. *Lecture Notes in Computer Science.* 2008; 5242:1–8.
92. Rahman M, Bhattacharya P, Desai BC. A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. *IEEE Trans Inf Technol Biomedicine.* 2007; 11:58–69.
93. Meurie C, Lebrun G, Lezoray O, Elmoataz A. A comparison of supervised pixels-based color image segmentation methods. Application in cancerology. *WSEAS Trans Computers.* 2003; 2:739–744.
94. Mao K, Zhao P, Tan P. Supervised learning-based cell image segmentation for P53 immunohistochemistry. *IEEE Trans Biomed Eng.* 2006; 53(5):1153–1163. [PubMed: 16761842]
95. Ranefalla P, Egevad L, Nordina B, Bengtsson E. A new method for segmentation of colour images applied to immunohistochemically stained cell nuclei. *Analytical Cellular Pathol.* 1997; 15:145–156.
96. Kothari S, Phan JH, Moffitt RA, Stokes TH, Hassberger SE, Chaudry Q, Young AN, Wang MD. Automatic batch-invariant color segmentation of histological cancer images. *Proc IEEE Int Symp Biomedical Imaging: From Nano to Macro.* 2011:657–660.
97. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Xiaojun G, Schmitt C, Thomas NE. A method for normalizing histology slides for quantitative analysis. *Proc IEEE Int Symp Biomedical Imaging: From Nano to Macro.* 2009:1107–1110.
98. Magee D, Treanor D, Crellin D, Shires M, Smith K, Mohee K, Quirke P. Colour normalisation in digital histopathology images. *Proc Optical Tissue Image Analysis Microscopy, Histopathology and Endoscopy (MICCAI Workshop).* 2009:100–111.
99. Mukherjee S, Roberts S. A theoretical analysis of the selection of differentially expressed genes. *J Bioinform Comput Biol.* 2005; 3:627–644. [PubMed: 16108087]
100. Chen L, Xuan J, Wang C, Shih I, Wang Y, Zhang Z, Hoffman E, Clarke R. Knowledge-guided multi-scale independent component analysis for biomarker identification. *BMC Bioinform.* 2008; 9
101. Phan J, Yin-Goen Q, Young A, Wang M. Improving the efficiency of biomarker identification using biological knowledge. *Proc Symp Biocomput.* 2009:427–438.
102. Tusher V, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Nat Acad Sci USA.* 2001; 98:5116–5121. [PubMed: 11309499]
103. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol.* 2005; 3:185–206. [PubMed: 15852500]
104. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S. Towards a knowledge-based human protein atlas. *Nature Biotechnol.* 2010; 28:1248–1250. [PubMed: 21139605]
105. Berglund L, Björling E, Oksvold P, Fagerberg L, Asplund A, Szigartyo CAK, Persson A, Ottosson J, Wernérus H, Nilsson P. A gene-centric human protein atlas for expression profiles based on antibodies. *Molecular Cellular Proteomics.* 2008; 7:2019–2027.
106. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma

- characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *CancerCell*. 2010; 17:98–110.
107. Bell D, Berchuck A, Birrer M, Chien J, Cramer D, Dao F, Dhir R, Disaia P, Gabra H, Glenn P. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011; 474:609–615. [PubMed: 21720365]
 108. Creighton CJ, Hernandez-Herrera A, Jacobsen A, Levine DA, Mankoo P, Schultz N, Du Y, Zhang Y, Larsson E, Sheridan R. Integrated analyses of microRNAs demonstrate their widespread influence on gene expression in high-grade serous ovarian carcinoma. *PloS One*. 2012; 7:e34546–e34546. [PubMed: 22479643]
 109. Shinjo K, Okamoto Y, An B, Yokoyama T, Takeuchi I, Fujii M, Osada H, Usami N, Hasegawa Y, Ito H. Integrated analysis of genetic and epigenetic alterations reveals CpG island methylator phenotype associated with distinct clinical characters of lung adenocarcinoma. *Carcinogenesis*. 2012
 110. Wang F, Kong J, Cooper L, Pan T, Kurc T, Chen W, Sharma A, Niedermayr C, Oh TW, Brat D. A data model and database for high-resolution pathology analytical image informatics. *J Pathology Informatics*. 2011; 2
 111. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform*. 2012
 112. Cheng Q, Chang JT, Geradts J, Neckers LM, Haystead T, Spector N, Lyster HK. Amplification and high-level expression of heat shock protein 90 marks aggressive phenotypes of human epidermal growth factor receptor 2 negative breast cancer. *Breast Cancer Res*. 2012; 14:R62–R62.
 113. Xiao Y, Decker PA, Rice T, McCoy LS, Smirnov I, Patoka J, Hansen HM, Wiemels JL, Tihan T, Prados MD. SSBP2 variants are associated with survival in glioblastoma patients. *Clin Cancer Res*. 2012
 114. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, Ladanyi M, Sander C. Integrative subtype discovery in glioblastoma using iCluster. *PloS One*. 2012; 7:e35236–e35236. [PubMed: 22539962]

Figure 2.
Cross-Laboratory Batch Effect in Gene Expression Data

Figure 3.
Histopathological Image Batch Effects

Figure 4.
Curse of Dimensionality: Exponential Growth of Sample-Space in High-Dimensional Problems

Figure 5.
Human Protein Atlas: Integration of Molecular and Cellular/Tissue Data

Figure 6.
Multi-Scale Data in the Cancer Genome Atlas

TABLE I**Renal Tumor Gene Expression Datasets**

Dataset	Platform	# Probes
Schuetz	Affy HG-Focus	8793
Jones	Affy HG-U133A	22283
Kort	Affy HG-U133+2.0	54675
Yusenko	Affy HG-U133+2.0	54675
Higgins	cDNA 2-Color	22689

TABLE II

Biomedical Knowledge Linked to Renal Tumor Subtypes

Renal Tumor Subtype	# of Microarray Probes Related to Cancer Gene Index Entries
Clear Cell	46
Chromophobe	12
Papillary	17
Oncocytoma	8

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE III

Biological Relevance of Ranking Renal Tumor Subtype Genes

Feature Ranking Method	Biological Relevance
Rank Products	0.69
Fold-Change	0.65
SAM	0.62
mRMR/D	0.58
mRMR/Q	0.58
T-test	0.58
Rank Sum	0.54