SAGE-Hindawi Access to Research
Human Genomics and Proteomics
Volume 2009, Article ID 869093, 13 pages
doi:10.4061/2009/869093

Review Article

Data Integration in Genetics and Genomics: Methods and Challenges

Jemila S. Hamid,[1] Pingzhao Hu,[2] Nicole M. Roslin,[2] Vicki Ling,[1,3] Celia M. T. Greenwood,[4] and Joseph Beyene[1,2,4]

[1] *Biostatistics Methodology Unit, The Hospital for Sick Children Research Institute, 555 University Avenue, Toronto, ON, Canada M5G 1X8*
[2] *The Center for Applied Genomics, The Hospital for Sick Children Research Institute, 555 University Avenue, Toronto, ON, Canada M5G 1X8*
[3] *Program in Developmental and Stem Cell Biology, The Hospital for Sick Children Research Institute, 555 University Avenue, Toronto, ON, Canada M5G 1X8*
[4] *Dalla Lana School of Public Health, University of Toronto, 555 University Avenue, Toronto, ON, Canada M5G 1X8*

Correspondence should be addressed to Joseph Beyene, joseph@utstat.toronto.edu

Received 25 September 2008; Accepted 1 December 2008

Recommended by Heikki Lehvaslaiho

Due to rapid technological advances, various types of genomic and proteomic data with different sizes, formats, and structures have become available. Among them are gene expression, single nucleotide polymorphism, copy number variation, and protein-protein/gene-gene interactions. Each of these distinct data types provides a different, partly independent and complementary, view of the whole genome. However, understanding functions of genes, proteins, and other aspects of the genome requires more information than provided by each of the datasets. Integrating data from different sources is, therefore, an important part of current research in genomics and proteomics. Data integration also plays important roles in combining clinical, environmental, and demographic data with high-throughput genomic data. Nevertheless, the concept of data integration is not well defined in the literature and it may mean different things to different researchers. In this paper, we first propose a conceptual framework for integrating genetic, genomic, and proteomic data. The framework captures fundamental aspects of data integration and is developed taking the key steps in genetic, genomic, and proteomic data fusion. Secondly, we provide a review of some of the most commonly used current methods and approaches for combining genomic data with focus on the statistical aspects.

1. Background

In recent years, increasing amounts of genomic data have become available. The size, type, and structure of these data have also been growing at an unprecedented rate. Gene expression, single nucleotide polymorphisms (SNP), copy number variation (CNV), proteomic, and protein-protein interactions are some examples of genomic and proteomic data produced using high throughput technologies such as microarrays [1], array comparative hybridization, aCGH [2], and mass spectrometry [3]. Each of these distinct data types provides a different, partly independent and complementary view of the whole genome. However, elucidation of gene function and other aspects of the genome may require more information than is provided by one type of data. The amount and type of biological data are expected to increase even further (e.g., methylation, alternative splicing, transcriptomic, metabolomic, etc.). This proliferation of experimental data makes systematic integration an important component of genomics and bioinformatics [4]. Data integration is increasingly becoming an essential tool to cope with the ever increasing amount of data, to cross-validate noisy data sets, and to gain broad interdisciplinary views of large genomic and proteomic data sets. Instances of combining and synthesizing data have increased considerably in the last several years and the need for improved and standardized methods has been recognized [5–7].

In functional genomics, for example, one is interested in defining the function of all the genes in the genome of an organism. Defining functions of genes is a daunting task and achieving this goal requires integrating information from different experiments [8]. Similarly, classifying a protein as a membrane protein based on protein sequences is a nontrivial task and has been the subject of much previous research, and it has been demonstrated that incorporating knowledge derived from aminoacid sequences, gene expression data, and known protein-protein interactions significantly improves classification performance compared to any single type of data [9].

The need for integration of heterogeneous data measured on the same individuals arises in a wide range of clinical applications as well. In this regard, the best example is perhaps the challenge that cancer researchers and clinicians face in the diagnosis, treatment, and prognostication of this complex disease. The clinical management of cancer is currently based for the most part on information accumulated from clinical studies [10]. However, cancer is thought to be primarily caused by random genetic alterations, and as such genomic data such as gene expression and protein data can be used to classify tumors into subtypes, and thus may have the potential to improve the clinical management of cancer.

Data integration also plays an important role in understanding environment-genome interactions in toxicogenomics, a discipline where one investigates how various genes respond to environmental toxins and stressors, and how these factors modify the function and expression of genes in the genome [11]. The contribution of different sources of data (such as genomics, proteomics, and SNP) in advancing the field of toxicology is discussed by Patel et al. [11].

Another more common type of data integration relates to combining similar types of data across different studies. This can, for example, be done through meta-analytic approaches. For instance, with the increasing number of publicly available independent microarray datasets, it is important to combine studies that address similar hypotheses. Such analyses have proven to be very useful in several applications (see, e.g., Rhodes et al. [12]).

There are a number of challenges in the context of high biology genetic and genomic studies. The challenges may be of conceptual, methodological, or practical nature and may relate to issues that arise due to experimental, computational, or statistical complexities. For example, genomic data are often subject to varying degrees of noise, the curse of high dimensionality, and small sample sizes. One may, therefore, benefit from integrating clinical, genomic, proteomic data along with environmental factors.

Integrating data from different sources brings many challenges. In dealing with heterogeneous data, for example, one needs to convert data from different sources into a common format and common dimension. Genomic data arises in the form of vectors, graphs, or sequences, therefore it is essential to carefully consider strategies that best capture the most information contained in each data type before combining them.

Moreover, data from different sources might have different quality and informativity. Probe design and experimental conditions are known to influence signal intensities and sensitivities for many high-throughput technologies [13, 14]. Even for similar data types, data from different sources might have different quality depending on the experimental conditions that generated the data. In microarray experiments, for instance, lack of standards generates heterogeneous datasets for which direct comparison and integration is not possible [15]. Data from different sources might also have different informativity even if their quality is good and reliable; thus one source of data might give us more information than the other in answering the biological question of interest. For example, gene expression microarray data is expected to provide more information in recognizing ribosomal proteins than protein-protein interaction data. However, expression data is not expected to provide much information in identifying membrane proteins [9].

The overarching goals of data integration are to obtain more precision, better accuracy, and greater statistical power than any individual dataset would provide. Moreover, integration can be useful in comparing, validating, and assessing results from different studies and datasets. It is likely that whenever information from multiple independent sources agree, it is more likely for the findings to be valid and reliable than information from a single source [8].

Current methods for data integration in general and combining genomic and genetic data in particular are scattered in the literature and lack solid conceptual framework. Putting them under a single framework would bring more understanding and clarity for the research community. With this background in mind, the objective of this paper is two fold. The first objective is to introduce a conceptual framework for integrating genomic and genetic data. This framework, which can be adapted to most data integration tasks in the life sciences, can serve as a guideline for understanding key issues and challenges that arise in data integration in genetics and genomics. We also believe that the framework we introduce here can be used for motivating and developing improved methods for integrating genomic and genetic data. The second purpose of the paper is to review some of the most commonly used current methods and approaches for combining genomic data. The reviews are done from a statistical perspective and our discussions are focused more on methodological issues and challenges. This could be useful in identifying research directions and might lead to improved methodologies in combining genomic data.

The paper is organized as follows. In Section 2, we provide a conceptual framework for data integration and discuss key concepts regarding this framework. In Section 3, we discuss some of the methods used in integrating similar data types, and methods for integrating heterogeneous data types, including integrating statistical results with biological data, are reviewed in Section 4. A brief discussion and some highlights for future research directions are presented in Section 5.

2. A Conceptual Framework for Data Integration

The concept of data integration is not well defined in the literature and it may mean different things for different people. For instance, Lu et al. defined data integration, in the context of functional genomics, as the process of statistically combining data from different sources to provide a unified view of the whole genome and make large-scale statistical inference [16]. We view data integration in a much broader context so that it includes not only combining of data using statistical approaches, but also data fusion with biological domain knowledge using a variety of bioinformatics and computational tools.

In this section, we propose a conceptual framework for integrating genomic and genetic data. This framework attempts to capture the fundamental aspects of data integration and is developed taking the key steps involved in genomic and genetic data fusion into consideration. A flowchart describing the conceptual framework is given in Figure 1. Below we briefly discuss each of the three key components of data integration: posing the statistical/biological problem; recognizing the data type; stage of integration.

2.1. Posing the Statistical/Biological Problem. Identifying the statistical or biological problem is the first step in any statistical research in general and in genomic and genetic data fusion in particular. Different directions in the framework and methods are followed depending on the biological question of interest. For example, one might merge preprocessed and transformed (independently or in parallel) microarray data from different labs (experiments) to increase sample size and answer a scientific question related to the detection of differentially expressed genes across a range of experimental conditions [17]. Traditional biological research questions are for the most part hypothesis-driven where one performs experiments to answer specific biological hypotheses. However, current high throughput data have a wealth of information in answering many other statistical or biological questions. In modern genomics, it is increasingly accepted to generate data in a relatively hypothesis-free setting where different questions can be posed on the pool of data and data are mined with a variety of computational and statistical tools with the hope of discovering new knowledge.

2.2. Data Types. Current data integration methods fall into two different categories—integrating similar data types (across studies) or integrating heterogeneous data types (across studies as well as within studies). Once we identify the biological or statistical question, we can ask ourselves what type of data we have. Classifying data as similar or heterogeneous is not an easy task. In this paper, we consider data as of "similar type" if they are from the same underlying source, that is, if they are all gene expression, SNP, protein, copy number, sequence, clinical, and so on. We refer to data as of the "heterogeneous type" if two or more fundamentally different data sources are involved. One might, for example, want to develop a predictive model based on different genomic data (SNP, gene expression, protein, sequence) as well as clinical data. These data sets might have different structures, dimensions, and formats. Some of them are sequences, some graphs, and yet others may be numerical quantities. Integration of heterogeneous data, therefore, entails converting each of the separate sources into common structure, format, and dimension before combining them.

Whether data are of similar or heterogeneous type, the issue of quality and informativity is of great importance as well. Each data source is subject to different noise levels depending on the technology, the platform, the lab, and many other systematic and random errors. Therefore, the concept of weighting the data sources with quality and/or informativity scores becomes an essential component of the framework.

2.3. Stages of Integration. Data from different sources can be integrated at three different stages—early, intermediate, or late. The stage at which data are combined depends on the biological question, the nature, and type of data as well as the availability of original data. Regardless of the biological question at hand (e.g., test for differential expression, class discovery, class prediction, gene mapping, etc.) one might, for example, merge data from different studies, experiments, or labs to increase sample size. This is considered as integration at early stage. Merging weighted (by quality and/or informativity scores) data is also considered as early integration. This is because attaching of weights to the data does not change the general format and nature of the resulting data. However, the integration is considered as intermediate if we transform individual data sources into another format before we combine them. For example, in class prediction problems, one might convert the data into similarity matrices such as the covariance or correlation matrix and combine these similarity matrices for better prediction. Unlike the early stage integration, original data sets from the different sources are converted to a common format and dimension. Integration is considered to be at a late stage if final statistical results from different studies are combined. This stage includes, among others, meta-analytic techniques where one typically combines effect sizes or p values across studies.

2.4. Preprocessing. Genomic data are subject to different noises and errors, and a number of critical steps are required to preprocess raw measurements. An important step considered in our framework, therefore, is preprocessing. However, this is not the main focus of this paper and hence we do not go into details. We refer the reader to [18, 19] for more details.

Preprocessing precedes data integration and may include background correction, normalization, and quality assessment of data from high throughput technologies [19]. Approaches for preprocessing vary depending on the type and nature of data. Preprocessing methods for microarray data are, for example, different from that for array CGH or proteomic data. Moreover, data from different technologies and platforms might be preprocessed differently.

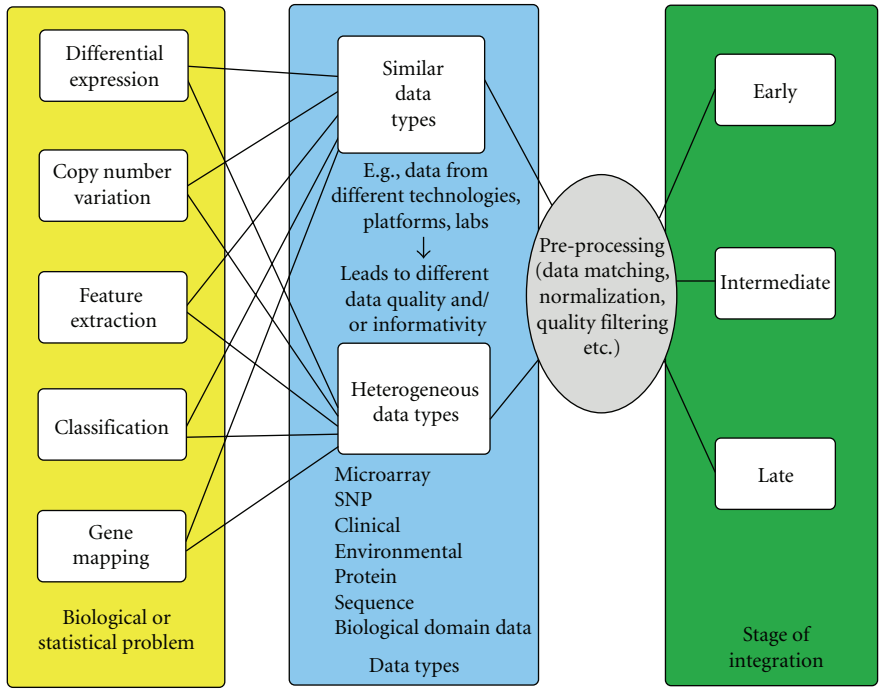

Figure 1: Conceptual framework for data integration in genetics and genomics.

For example, different approaches are utilized for preprocessing cDNA and Affymetrix gene expression microarray data [18, 19].

Data preprocessing can be done at any step of the data integration process. Some form of preprocessing is almost always done at the initial stage [13]. However, genomic data, in most cases, has to go through some sort of preprocessing before doing any statistical analysis to answer the biological question of interest. There is a large body of literature on this topic, and several ways of preprocessing high throughput data have been proposed [18, 19]. Approaches, both graphical and statistical, are also available for visualizing and checking if data needs to be preprocessed. In microarray studies, the standard procedure for researchers is to use preprocessed data as the starting point; however, this has prevented many researchers from carefully considering their preprocessing methods [20].

Data matching is another preprocessing step that needs to be taken into account before combining data from high throughput technologies. In gene or marker specific data integration, a major challenge in pooling information from different data sources can be partly due to the fact that measurements are obtained by different technologies or they measure different aspects of the same underlying quantity. For example, Affymetrix uses different numbers of probes to measure the same gene in different chip types. Therefore, it is impossible to get comparable gene expression levels across different chip types simply based on gene identifiers since they used different probe sequences for the given target probes. Mecham et al. [21] proposed a sequence-based matching of probes instead of gene identifier-based matching. The results showed that at different levels of the analysis (e.g., gene expression ratios), cross-platform consistency is significantly improved by sequence-based matching.

Combining data from different genotyping projects presents a problem in that different marker sets are used for different arrays. For example, out of the approximately 1 million SNPs on each of the Affymetrix Human SNP 6.0 and Illumina Human1M arrays, only about 250,000 SNPs are common to both assays [22]. To overcome this difficulty, genotype imputation algorithms, such as MACH [23], IMPUTE [24], and fastPHASE [25], have been developed to impute alleles at ungenotyped markers, based on the genotypes of surrounding markers. In summary, the preprocessing step can hugely affect the properties of the final statistical summaries and hence the statistical results. Therefore, methods for preprocessing must be chosen with care.

3. Integrating Similar Data Types

Although posing the statistical/biological problem is the first step in any study involving data integration, data integration in general is divided into two broad categories: integration of similar data types and integration of heterogeneous data types. In this section, we give a review of some of the current methods available for integrating similar data types. Similar data types have been combined to answer different biological questions, and approaches for integrating such data at early, intermediate, or late stages have been proposed in literature. However, these approaches are in general meta-analytic methods. Some examples from the published literature that can serve as illustration for data integration concepts

corresponding to different biological/statistical questions, data types, and stages of integration are highlighted in Table 1.

3.1. Integration of Linkage Studies. Linkage analysis is a gene mapping technique which is based on the process of recombination, or the crossing over of parental chromosomes when forming gametes which will eventually be passed on to offspring. In order to observe or infer recombination events, families are required to reconstruct the transmission of alleles and phenotypes through several generations. Highly polymorphic markers are genotyped either in a region of interest, based on previous knowledge or hypotheses (candidate gene mapping), or across the entire genome (genome-wide scans). Linkage analysis looks for the cosegregation of a marker and the trait of interest in each family, and one way to assess linkage statistically is through the log odds, or LOD score:

$$\text{LOD} = \log 10(L(\text{data}|\text{theta})/L(\text{data}|\text{theta} = 0.5)),$$

where L is a likelihood function and theta is a measure of the amount of recombination between the marker and the trait. The LOD score can be maximized over a grid of theta values, but is usually just computed at fixed theta intervals. Scores for each family are summed, and an overall LOD score > 3 (i.e., likelihood ratio than 1000:1 odds) is generally considered to be a significant evidence for linkage. For further details, the interested reader is referred to an excellent textbook in [30].

Linkage analysis has been very successful for rare Mendelian disorders, that is, diseases where the variant is associated with a large increase in risk, and usually only one variant is responsible for the phenotype. However, in the case of complex traits, where multiple variants are likely to contribute, each with more modest risk, large collections of families need to be genotyped and phenotyped to have modest power to detect linkage, which is a costly and time-consuming undertaking. Combining information from several scans can, therefore, help overcome these difficulties. Three main strategies have been used to integrate linkage data: pooling of datasets (integration at an early stage), combining linkage statistics or P-values (integration at a late stage), and combining effect sizes (also integration at a late stage).

When the raw data from all studies are available, the most powerful approach is to simply pool the datasets and analyze this large dataset as if it were one study, termed a "mega analysis" [31]. This assumes that the same markers are common to all studies, and that there were identical ascertainment strategies and similar allele frequencies in the populations from which the samples were derived from. It is also equivalent to simply sum the LOD scores from various studies, under the same restrictions [32]. Since it is rare to have identical marker maps across studies, methods to allow LOD score calculations at arbitrary marker positions were developed to overcome this problem [33, 34]. An alternative method, known as the posterior probability of linkage (PPL), puts the question of linkage in a Bayesian context, which allows the posterior distribution of linkage, given the data, to be updated as new data and studies are accumulated [35].

In the more common situation when the raw data is not available, perhaps the simplest method to combine independent P-values is the one developed by Fisher [36],

$$S = -2\sum_{i=1}^{k} \log p_i, \tag{1}$$

where k is the number of studies, p_i is the P-value obtained from study i. The statistic S is asymptotically distributed as a chi-square with $2k$ degrees of freedom. This procedure was used by Allison and Heo [37] to identify regions linked to obesity, and a modified version was used by Badner and Gershon [38] for autism studies. The null hypothesis for this test is that none of the studies show significant results, against the alternative that at least one is significant. This may not be the question that a researcher is interested in asking; a more relevant question may be whether all studies support a common hypothesis. In this situation, Rice (1990) [39] suggests that a summary statistic based on the mean of the normal-transformed P-values may be more appropriate.

In the Fisher method, a single P-value per study is used. However, in the context of a genome-wide scan, many markers are genotyped, and so many LOD scores or P-values are calculated. To address this issue, Wise et al. [40] proposed a method called genome search meta-analysis (GSMA) which ranks significant results within each study, and sums the ranks for each region across all studies. The ranks can be weighted by study characteristics such as the number of pedigrees or the number of markers. This test will detect regions which are implicated in several studies.

The use of P-values from individual studies generally precludes the estimation of average effect sizes, which can be of interest in linkage studies. These estimates are the main goal of most standard meta-analyses. Li and Rao [41] proposed the use of a random effects model [42] to combine regression coefficients from a linear model of the squared trait differences as the dependent variable, and the proportion of alleles shared identical by descent (IBD) at a marker for sibling pairs as the independent variable. The model was also applied using the proportion of alleles shared IBD directly as the measure of effect size [43]. Along with an overall estimate of effect size, the random effects model has the advantage of being able to test for and control for differences in effect sizes across studies (sometimes referred to as heterogeneity).

3.2. Integration of Genetic Association Studies. In the past few years, it has been shown that genome-wide association studies have strong power to identify genetic determinants for common and rare diseases. Due to the high cost of performing these types of studies, it becomes more and more important to integrate evidence from multiple studies in characterizing the genetic risks of these diseases. Meta-analyses can offer both enhanced power to detect associations and increased precision of estimates of its magnitude. There are two major methods with focusing on late stage integration. One is combining effect sizes, primarily the odds ratio

Table 1: Some illustrative examples for integrating similar and heterogeneous genomic, genetic, and proteomic data.

Data types	Biological/statistical question	Stages of integration	Example/comments
Similar data types	Sample classification	Early	Jiang et al. [17] integrated two Affymetrix data sets; each data set was first distribution transformed and two data sets were then merged together.
	Differential gene analysis	Late	Rhodes et al. 2002 [12] integrated two cDNA and two Oligo data sets; *P*-value was first calculated for each gene in each study and Fisher's method was used to combine the *P*-values.
	Gene mapping	Late	Ioannidis et al. [26] integrated two Affymetrix and one Illumina SNP data sets; odds ratio (OR) was first calculated for each SNP in each study and random effects model was used to combine ORs.
Heterogeneous data types	Candidate gene discovery	Intermediate	Adler et al. [27] integrated aCGH and gene expression data sets; association analysis between two datasets was made in amplification and deletion regions.
	Protein classification	Intermediate	Lanckriet et al. [9] integrated sequence, interaction, and expression; a kernel matrix was first generated for each data set and combined using optimal weights.
	Gene mapping	Intermediate	McCaroll and Altshuler (2007) [28] integrated genotype and copy number variation data; copy number variation as covariates was used in association analysis.
	Gene set (function) differential analysis	Intermediate	Al-Shahrour et al. [29] integrated gene expression and biological domain information; gene-specific test statistic was first calculated, which was then integrated with biological domain information to evaluate function enrichment.

(OR) and another is to combine *P*-values [26, 44, 45]. The effect size based method can be fixed effects or random effects models. For example, Ioannidis et al. [26] applied a random effects model to combine all data sets generated in three stages from three genome-wide association (GWA) studies on type 2 diabetes. Details of the design and populations of these studies have been presented in the original publications [46–48]. Ioannidis et al. [26] selected 11 polymorphisms suggested as susceptibility loci for type 2 diabetes in at least one of the three studies. They found 5 of the 11 genetic variants have moderate to very large heterogeneity across studies. Therefore, they used random effects calculations incorporating between study heterogeneity for these 5 polymorphisms and found more conservative *P*-values for the summary effects compared with the fixed effects calculations. Instead of focusing on meta-analysis of only identified polymorphisms, Zeggini et al. [45] applied fixed effects model and combing *P*-value methods to meta-analyze the same data sets, they detected at least six previously unknown loci with robust evidence for association.

3.3. Integration of Gene Expression Microarray Studies. Microarrays have been widely used in identifying differentially expressed genes [48, 49] and for building gene expression profile-based predictors for disease outcome diagnosis [50–54]. Although some of these studies have led to promising results [51], it is difficult to directly compare the results obtained by different groups addressing the same biological problem. This is because laboratory protocols, microarray platforms, and analysis techniques used in each study may not be identical [55, 56]. Moreover, most individual studies have relatively small sample sizes, and hence

predictive models trained by individual studies using cross-validation are prone to over-fitting, leading to prediction accuracies that may be less robust and lack generalizability [57]. Recent studies show that systematic integration of gene expression data from different sources can increase statistical power in detecting differentially expressed genes while allowing for an assessment of heterogeneity, and may lead to more robust, reproducible, and accurate predictions [12, 15, 17, 56, 58–62]. Therefore, our ability to develop powerful statistical methods for efficiently integrating related genomic experiments is critical to the success of the massive investment made on genomic studies. Here, we highlight some of the strategies that have been used to integrate microarry gene expression studies.

Combining gene expression data at early and late stages has been considered by different groups. In integrating gene expression data at an early stage, data sets generated in each study are first preprocessed independently or in parallel, and then the preprocessed datasets are put together so that the integrated data set can be treated as one data set. In this way, the sample size of the study is greatly increased. Several transformation methods have been proposed to process gene expression measures from different studies [17, 56, 59, 62]. For example, Jiang et al. [17] transformed the normalized data sets to have similar distributions and then merged the transformed data sets. Wang et al. [59] standardized each gene expression level based on the average expression measurements and the standard errors estimated from prostate cancer samples. These methods are simple and in many cases, if the transformation is carefully made, lead to improved prediction [17]. Nevertheless, there are no consensus or clear guidelines as to the best way to perform such data transformations.

At the late stage, results from statistical analyses are combined using meta-analytic approaches. Similar to the case of linkage and association gene mapping studies, one of the popular approaches combines effect sizes from different studies while taking interstudy variability into account when estimating the overall mean for each gene across studies. For example, Choi et al. [15] focused on integrating effect size estimates in individual studies into an overall estimate of the average effect size. The effect size was used to measure the magnitude of treatment effect in a given study and random effects model was adopted to incorporate interstudy variability. Using the same microarray data sets as those used by Rhodes et al. [12], Choi et al. [15] demonstrated that their method can lead to the discovery of small but consistent expression changes with increased sensitivity and reliability among the datasets. For each gene, the widely used effect size measure is the standardized mean difference which is obtained by dividing the difference in average gene expression between groups of interest by a pooled estimate of standard deviation [63, 64]. It is well known in microarray data analysis that the estimated standard deviation might be unstable when the sample size in each group is small. Therefore, much effort has been made to overcome the shortfall by using a penalty parameter for smoothing the estimates using information from all genes rather than relying solely on the estimates from an individual gene [4, 65].

As mentioned before, the other meta-analytic technique commonly used combines *P*-values across different studies. For example, Rhodes et al. [12, 66] integrated results from prostate cancer microarray studies which have been performed on different platforms. Differential expression was first assessed independently for each gene in each dataset using *P*-values and *P*-values from individual studies were combined using Fisher's method (see also Section 3.1). Their analysis revealed stronger evidence for statistical significance from the combined analysis than any of the individual studies separately. Combining *P*-values can be useful in detecting effects with improved statistical significance, but this method does not indicate the direction of significance (e.g., up- or downregulation) [67]. Instead of integrating *P*-values directly, some studies explored combining the ranks of the *P*-values from different studies [61, 68]. For example, DeConde et al. [61] proposed a rank-aggregation method and combined microarray results from five prostate cancer studies where they showed that their approach can identify more robust differentially expressed genes across studies.

The data integration approaches discussed above to integrate microarrays are in a quality-unweighted framework [12, 15, 17, 56, 58, 59, 61, 62, 66]. However, it has been argued that studies of higher quality give more accurate estimates and, as a result, should receive higher weight in the analysis summarizing findings across studies [69]. In gene expression microarrays, many genes may be "off" or not detectable in a particular adult tissue, moreover, some genes may be poorly measured due to probes that are not sufficiently sensitive or specific. Therefore, the signal strength and clarity will vary across the genes, suggesting that a quality measurement could highlight strong and clear signals [70, 71]. How to best measure the quality of a gene expression measurement and how best to use such a quality measure are still open research questions. However, different strategies can be considered for incorporating quality weights into meta-analysis of microarray studies. For example, a quality threshold can be defined and only genes that are above this threshold can be included in the meta-analysis. However, the choice of threshold will be arbitrary. In a recent study, our group proposed a quality measure based on the detection *P*-values estimated from Affymetrix microarray raw data [60, 70]. Using an effect-size model, we demonstrated that the incorporation of quality weights into the study-specific test statistics, within a meta-analysis of two Affymetrix microarray studies, produced more biologically meaningful results than the unweighted analysis.

4. Integrating Heterogeneous Data Types

Perhaps the most challenging type of data integration is combining heterogeneous data types. A wide variety of genomic and proteomic data are becoming available at an unprecedented pace including data on, but not limited to, gene expression (quantitative numbers), gene/protein sequences (strings), gene-gene/protein-protein interactions (graphs). There is also a growing interest for integration

of these and related molecular information with clinical, laboratory, as well as environmental data. Broadly speaking, integrating heterogeneous data types involves two steps. The first one is converting data from different sources into a common format. The second, equally important, step is to combine the data and perform statistical analysis on the combined data set. Here we survey some of the currently available approaches for integrating heterogeneous data types. Some illustrative examples are highlighted in Table 1. An illustrative flowchart outlining integrative analyses of heterogeneous data for finding disease-causing genes is shown in Figure 2.

4.1. Integration of Gene Expression with Genotype Data. Gene expression levels of many genes have been successfully used to show natural variation in humans [72, 73]. Using regression analysis where the dependent variables are expression levels and the independent variables are the genotypes, it has been shown that expression levels may be influenced by single nucleotide polymorphisms [72–75]. These mapping efforts have identified quantitative trait loci (QTLs) that may be in the gene's own regulatory regions (*cis*-acting QTLs) as well as elsewhere in the genome (*trans*-acting QTLs) using both linkage [72] and association analysis [73, 74]. For the association analysis, Stranger et al. [74] examined all possible combinations of gene expression phenotype/marker genotype combinations, whereas Cheung et al. [73] examined only gene expression phenotype/genotype combinations under linkage peaks identified in the study by Morley et al. [72]. However, it may be possible that multiple loci play a role in regulating the expression level of a single phenotype. To this effect, our group used the gene expression and SNP data reported in Morley et al. [72] and applied stepwise regression analysis to look for additive effects of the SNPs which led to the identification of *cis*- and *trans*-acting loci that regulate gene expression [75]. We identified many expression phenotypes that have significant evidence of association and linkage to one or more chromosomal regions and showed that much of the observable heritability in the phenotypes could be explained by simple SNP associations.

Due to the large number of genes in current high volume data sets and the existence of various degrees of noise in the data, integration involving all single-nucleotide polymorphism (SNP) loci and gene expression phenotypes may computationally challenging, and results may lack biological plausibility and interpretability. One promising approach, that is computationally efficient and can lead to more robust and interpretable results, is to use methods that induce sparseness in the integrated solutions where noisy data are automatically filtered out from analysis. For example, our group has recently introduced a novel sparse canonical correlation analysis (SCCA) statistical method, which allowed us to examine the relationships of many genetic loci and gene expression phenotypes by providing sparse linear combinations that include only a small subset of loci and gene expression phenotypes [76]. The correlated sets of variables resulting from sparse canonical correlation analysis are sufficiently small for biological interpretability and further follow up. We applied SCCA to data reported in [72] and identified small but interesting group of loci and gene expressions that have maximal correlation across the two data sources (gene expression and genotypes).

4.2. Integration of Copy Number Variation and Gene Expression Data. Array CGH (aCGH) microarray technology has been widely used to identify small or large regions of amplifications and deletions along the genome of organisms. Recent studies have tried to incorporate gene expression data with aCGH data for finding disease causing genes [27, 77–79]. For example, Pollack et al. [77] analyzed gene expression levels in parallel with copy number aberrations for the same set of breast tumors. They found that DNA copy number does have an impact on gene expression levels, and that a 2-fold change in DNA copy number corresponds to an average 1.5-fold change in expression level [72–75]. However, it has also been observed that many overexpressed genes were not amplified and that not all amplified genes were highly expressed, but the small number of genes that were overexpressed and amplified could be interesting genes. For example, Platzer et al. [80] measured gene expression levels and DNA copy numbers of colon cancer samples, and found four chromosomal arms that contained amplifications in most samples. Among expression levels of 2146 transcripts on these arms, only 81 have greater than 2-fold change in gene expression. They concluded that chromosomal amplifications do not result in global over expression of the genes located at that position. Huang et al. [79] also found that genomic DNA copy number aberrations (amplification or deletion) appeared not to be parallel with the corresponding gene expressions in any given samples. Most of these methods explore the relationship of genomic DNA copy number and gene expression at relatively the same positions of the genes on the genome. However, since it is known that genes on a chromosome are coregulated [81], a better way is to determine clusters of significantly over or under expressed genes by taking the chromosome position into account. This can also be applied on CGH data, and then a correlation of the clustered or coregulated expression signatures and copy number data can be determined [78].

4.3. Kernel Based Data Integration for Class Prediction. One of the novel and most promising methods for integrating heterogeneous data types are kernel-based statistical methods. In kernel-based methods, each data set is represented by a so-called kernel matrix, which essentially constitutes similarity measures between pairs of entities (genes, proteins, patients, etc.). In general, these methods are applied in biological problems related to class discovery and class prediction. In functional genomics, for example, one is interested in discovering new functional classes and/or assigning each gene or protein into already existing classes. It is also useful in cancer research where one is interested in discovering new tumor subtypes and/or assigning patients into already existing tumor types. Kernel-based statistical methods are tools which have already proven to be powerful in areas such as pattern recognition, chemoinformatics,

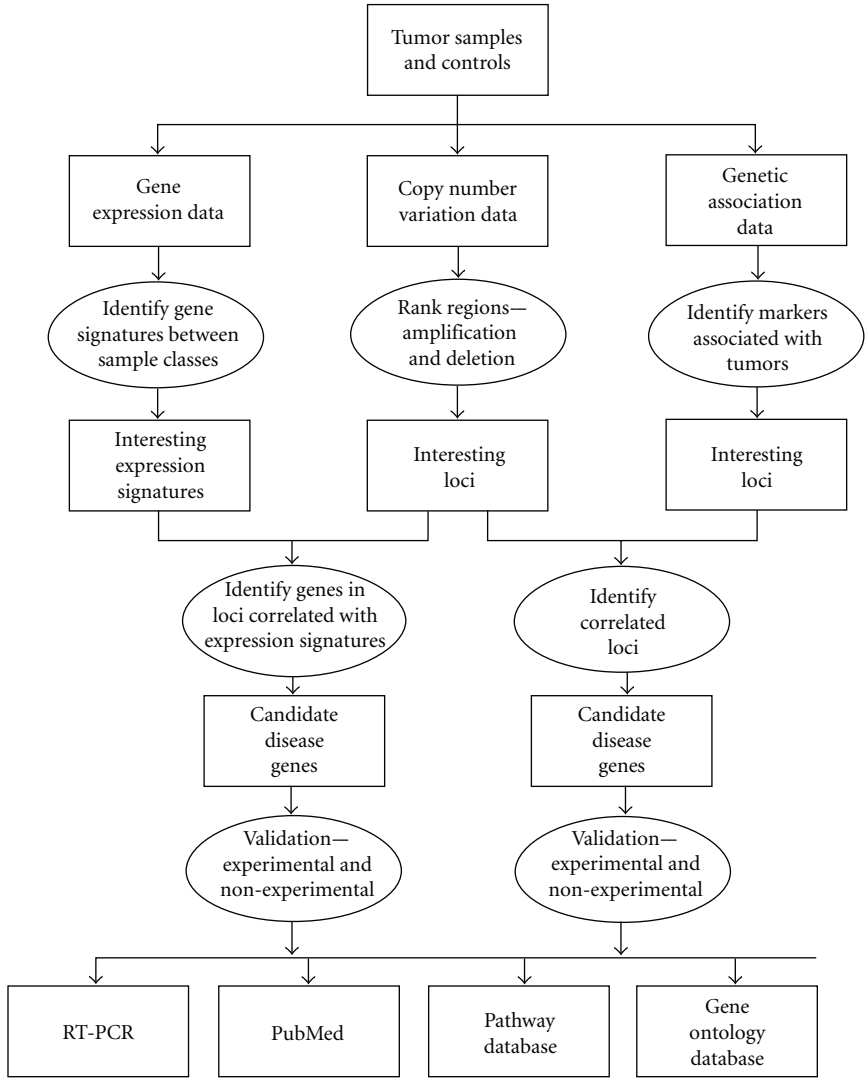

Figure 2: An illustrative flowchart for finding disease causing genes by integrating heterogeneous data.

computational linguistics, and bioinformatics. Their rapid uptake in these applications is due to their reliability, accuracy, and computational efficiency as well as their ability to handle various types of data [82].

One can describe kernel-based statistical learning approaches using two basic steps. The first one is choosing the right kernel for each data set. This is a crucial and difficult step in combining heterogeneous data using kernels. One of the reasons for the success of kernel methods is that kernel matrices can be defined for any type of data as well as their ability to incorporate prior knowledge through the kernel function [82]. The choice of kernel matrices depends on what type of data we have (e.g., diffusion kernel for graphical data and sequence-based kernels based on algorithms such as BLAST for protein sequences) and if the patterns in the data are linear (linear or standardized linear kernel can be used) or nonlinear (Gaussian or other nonlinear kernels can be used). One can also define different kernels on the same data sets. This allows us to get different view of the same data set and might provide us more information than using a single kernel. For example, Lanckriet et al. [9] defined 7 different kernels on three different data sources for predicting membrane and ribosomal proteins. The second equally important and challenging step deals with combining the kernels from the different data sources to give a complete representation of available data for a given statistical task. Basic mathematical operations such as multiplication, addition, and exponentiation preserve properties of kernel matrices and hence produce valid kernels. The simplest approach is to use the sum of the different kernels, which is equivalent to taking the average kernel. This naïve combination has been used mainly for comparison purposes. However, not all data have equal quality and informativity. Depending on the statistical and biological question at hand, data from one source might contain more information than the other. Moreover, the quality of data might vary because of different limitations and factors involved in different experiments. To our knowledge, currently there are no

published methods that explicitly incorporate quality and informativity measures into the kernel framework.

There are few kernel-based statistical methods proposed for integrating heterogeneous genomic data. Lanckriet et al. [9] used kernel-based support vector machine (SVM) method to recognize particular classes of proteins—membrane proteins and ribosomal proteins. Their method finds the classification rule as well as the corresponding weights for each data set. The performance of the SVM trained on the combined data set is better than that of the SVM trained on each of the individual data sets. Moreover, the weights produced from the algorithm give some measure of the relative importance of the different data sets. Another similar kernel-based approach was used by Daemen et al. [4, 10]. They used kernel-based least square support vector machine (LS-SVM) to combine clinical and microarray data [10]. The same group applied their method to combine microarray and proteomics data [4]. They chose a standardized linear kernel for both data sets in both papers. In the first paper, leave-one-out cross validation was performed on the training data set to get optimal weights. The model based on the clinical and microarray data performed slightly better than the model based on each of the data sets alone. The performance of their method was also compared with three conventional clinical prognostic indices and was shown that the kernel-based integrated microarray and clinical data outperforms all three conventional approaches. In the second study, the authors used the same method to combine microarray and proteomic data to predict the response on cetuximab in patients with rectal cancer. Tissue and plasma samples were taken from the patients before treatment and at the moment of surgery. Tissues were used for microarray analysis and plasma samples were used for proteomics analysis. They defined four kernels from these data sets and assigned equal weights to each one of them, that is, a naïve combination of kernels was used. The method trained on microarray data (with 5 genes) and protein data (10 proteins) performed better than any of the other alternatives they considered.

4.4. Integrating Statistical Results with Biological Domain Data. The ultimate purpose of statistical analysis on genomic data is to gain some insight into the fundamental biology. Annotation of *statistical* results helps biologists in interpreting discovered patterns. A wide variety of biological information is available to the public, such as information on published literature on the topic of interest (e.g., PubMed) and functional/pathway information (e.g., Gene Ontology, KEGG). Integrating biological information with statistical results is, therefore, another important type of data integration which can be considered as a bridge between statistical results and biological interpretation. Including biological domain data in statistical analysis can be done at any stage of analysis. Al-Shahrour et al. [29], for example, combined statistical results from gene expression data with biological information in discovering molecular functions related to certain phenotypes. Another popular approach to incorporate prior biological knowledge into statistical analysis is gene set enrichment analysis (GSEA) [83, 84]. Given an a priori defined set of genes, the goal of GSEA is to determine whether a particular gene is enriched or not, that is, whether members are randomly distributed throughout or primarily found at the top or bottom of a ranked list of gene differential expression results.

There is also a rapidly growing list of computational and visualization tools that can be used to integrate statistical findings with biological domain information, and thereby facilitating interpretation. For instance, packages from the bioconductor project (www.bioconductor.org) provide powerful analytical annotation and visualization tools for a wide range of genetic and genomic data sets.

5. Summary and Future Directions

With a rapidly increasing amount of genomic, proteomic, and other high throughput data, the importance of data integration has increased significantly. Biologists, medical scientists, and clinicians are also interested to integrate recently available high throughput data with already existing clinical, laboratory, as well as prior biological information. Moreover, data have been produced in various formats (graphs, sequences, vectors, etc.) and dimensions and, as a result, a simple merge of available data is not applicable and in some cases impossible. Furthermore, data from different sources are subject to different noise levels due to difference in technologies, platforms and other systematic or random factors affecting the experiments. Consequently, data might have different qualities and a naïve combination of data is not appropriate in such cases. The concept of data informativity is also essential in any data integration problem. Data from various sources might contain different informativity for a given statistical or biological task. One data source might, for example, be more informative than the other. A good data integration method should, therefore, take these into account. Even if quality scoring has been used in traditional statistical analysis, use of quality weights is not common in genetics and genomics. Moreover, appropriate quality and informativity measures have not been defined for many data types. An extensive research is, therefore, needed in developing quality and informativity scores for various genomic, genetic, and proteomic data.

In this paper, we proposed a conceptual framework for genomic and genetic data integration. This framework, with a little modification, can also be useful in any data integration problem. The framework provides different steps involved in genomic data integration and addresses different issues and challenges. Moreover, putting current methodologies for data integration under a single framework brings more understanding in the research community. Furthermore, we hope that it would play an important role in the development of standardized and improved data integration methods that takes the quality, informativity, and other aspects of individual data sets.

Acknowledgments

This work was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Mathematics of Information Technology and Complex Systems (MITACS), the Canadian Institute of Health Research (CIHR) (grant number 84392), and Genome Canada through the Ontario Genomics Institute. The auhtors would also like to thank two anonymous reviewers for helpful comments.

References

[1] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[2] A. E. Oostlander, G. A. Meijer, and B. Ylstra, "Microarray-based comparative genomic hybridization and its applications in human genetics," *Clinical Genetics*, vol. 66, no. 6, pp. 488–495, 2004.

[3] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.

[4] A. Daemen, O. Gevaert, T. De Bie, et al., "Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer," *Pacific Symposium on Biocomputing*, vol. 13, pp. 166–177, 2008.

[5] D. M. Reif, B. C. White, and J. H. Moore, "Integrated analysis of genetic, genomic and proteomic data," *Expert Review of Proteomics*, vol. 1, no. 1, pp. 67–75, 2004.

[6] T. C. Prevost, K. R. Abrams, and D. R. Jones, "Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening," *Statistics in Medicine*, vol. 19, no. 24, pp. 3359–3376, 2000.

[7] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein, "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 14, pp. 8348–8353, 2003.

[8] R. Jansen, N. Lan, J. Qian, and M. Gerstein, "Integration of genomic datasets to predict protein complexes in yeast," *Journal of Structural and Functional Genomics*, vol. 2, no. 2, pp. 71–81, 2002.

[9] G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and S. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, no. 16, pp. 2626–2635, 2004.

[10] A. Daemen, O. Gevaert, and B. De Moor, "Integration of clinical and microarray data with kernel methods," in *Proceedings of the 29th Annual International Conference of IEEE Engineering in Medicine and Biology Society (EMBC '07)*, pp. 5411–5415, Lyon, France, August 2007.

[11] S. Patel, D. Parmar, Y. K. Gupta, and M. P. Singh, "Contribution of genomics, proteomics, and single-nucleotide polymorphism in toxicology research and Indian scenario," *Indian Journal of Human Genetics*, vol. 11, no. 2, pp. 61–75, 2005.

[12] D. R. Rhodes, J. Yu, K. Shanker, et al., "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 25, pp. 9309–9314, 2004.

[13] "Affymetrix Microarray Suite User Guide," Version 5, 2001.

[14] J. Listgarten and A. Emili, "Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry," *Molecular & Cellular Proteomics*, vol. 4, no. 4, pp. 419–434, 2005.

[15] J. K. Choi, U. Yu, S. Kim, and O. J. Yoo, "Combining multiple microarray studies and modeling interstudy variation," *Bioinformatics*, vol. 19, supplement 1, pp. i84–i90, 2003.

[16] L. J. Lu, Y. Xia, A. Paccanaro, H. Yu, and M. Gerstein, "Assessing the limits of genomic data integration for predicting protein networks," *Genome Research*, vol. 15, no. 7, pp. 945–953, 2005.

[17] H. Jiang, Y. Deng, H.-S. Chen, et al., "Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes," *BMC Bioinformatics*, vol. 5, article 81, pp. 1–12, 2004.

[18] Y. H. Yang, S. Dudoit, P. Luu, et al., "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Research*, vol. 30, no. 4, p. e15, 2002.

[19] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Research*, vol. 31, no. 4, p. e15, 2003.

[20] Z. Wu and R. A. Irizarry, "Preprocessing of oligonucleotide array data," *Nature Biotechnology*, vol. 22, no. 6, pp. 656–658, 2004.

[21] B. H. Mecham, G. T. Klus, J. Strovel, et al., "Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements," *Nucleic Acids Research*, vol. 32, no. 9, p. e74, 2004.

[22] P. I. W. de Bakker, M. A. R. Ferreira, X. Jia, B. M. Neale, S. Raychaudhuri, and B. F. Voight, "Practical aspects of imputation-driven meta-analysis of genome-wide association studies," *Human Molecular Genetics*, vol. 17, no. R2, pp. R122–R128, 2008.

[23] Y. Li and G. R. Abecasis, "Mach 1.0: rapid haplotype reconstruction and missing genotype inference," *American Journal of Human Genetics*, vol. S79, p. 2290, 2006.

[24] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, "A new multipoint method for genome-wide association studies by imputation of genotypes," *Nature Genetics*, vol. 39, no. 7, pp. 906–913, 2007.

[25] P. Scheet and M. Stephens, "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase," *American Journal of Human Genetics*, vol. 78, no. 4, pp. 629–644, 2006.

[26] J. P. A. Ioannidis, N. A. Patsopoulos, and E. Evangelou, "Heterogeneity in meta-analyses of genome-wide association investigations," *PLoS ONE*, vol. 2, no. 9, p. e841, 2007.

[27] A. S. Adler, M. Lin, H. Horlings, D. S. A. Nuyten, M. J. van de Vijver, and H. Y. Chang, "Genetic regulators of large-scale transcriptional signatures in cancer," *Nature Genetics*, vol. 38, no. 4, pp. 421–430, 2006.

[28] S.A. McCaroll and D. Altshuler, " Copy-number variation and association studies of human disease," *Nature Genetics*, vol. 39, pp. 37–42, 2007.

[29] F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo, "Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information," *Bioinformatics*, vol. 21, no. 13, pp. 2988–2993, 2005.

[30] J. Ott, *Analysis of Human Genetic Linkage*, Johns Hopkins University Press, Baltimore, Md, USA, 3rd edition, 1999.

[31] E. Lander and L. Kruglyak, "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results," *Nature Genetics*, vol. 11, no. 3, pp. 241–247, 1995.

[32] N. E. Morton, "Sequential tests for the detection of linkage," *American Journal of Human Genetics*, vol. 7, no. 3, pp. 277–318, 1955.

[33] E. S. Lander and P. Green, "Construction of multilocus genetic linkage maps in humans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 84, no. 8, pp. 2363–2367, 1987.

[34] D. W. Fulker, S. S. Cherny, and L. R. Cardon, "Multipoint interval mapping of quantitative trait loci, using sib pairs," *American Journal of Human Genetics*, vol. 56, no. 5, pp. 1224–1233, 1995.

[35] V. J. Vieland, "Bayesian linkage analysis, or: how I learned to stop worrying and love the posterior probability of linkage," *American Journal of Human Genetics*, vol. 63, no. 4, pp. 947–954, 1998.

[36] R. A. Fisher, *Statistical Methods for Research Workers*, Hafner, New York, NY, USA, 12th edition, 1954.

[37] D. B. Allison and M. Heo, "Meta-analysis of linkage data under worst-case conditions: a demonstration using the human OB region," *Genetics*, vol. 148, no. 2, pp. 859–865, 1998.

[38] J. A. Badner and E. S. Gershon, "Regional meta-analysis of published data supports linkage of autism with markers on chromosome 7," *Molecular Psychiatry*, vol. 7, no. 1, pp. 56–66, 2002.

[39] W.R. Rice, " A consensus combined *P* -value test and the family-wide significance of component tests," *Biometrics*, vol. 46, pp. 303–308, 1990.

[40] L. H. Wise, J. S. Lanchbury, and C. M. Lewis, "Meta-analysis of genome searches," *Annals of Human Genetics*, vol. 63, no. 3, pp. 263–272, 1999.

[41] Z. Li and D. C. Rao, "Random effects model for meta-analysis of multiple quantitative sibpair linkage studies," *Genetic Epidemiology*, vol. 13, no. 4, pp. 377–383, 1996.

[42] R. DerSimonian and N. Laird, "Meta-analysis in clinical trials," *Controlled Clinical Trials*, vol. 7, no. 3, pp. 177–188, 1986.

[43] C. Gu, M. Province, A. Todorov, and D. C. Rao, "Meta-analysis methodology for combining non-parametric sibpair linkage results: genetic homogeneity and identical markers," *Genetic Epidemiology*, vol. 15, no. 6, pp. 609–626, 1998.

[44] K. E. Lohmueller, C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn, "Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease," *Nature Genetics*, vol. 33, no. 2, pp. 177–182, 2003.

[45] E. Zeggini, M. N. Weedon, C. M. Lindgren, et al., "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes," *Science*, vol. 316, no. 5829, pp. 1336–1341, 2007.

[46] L. J. Scott, K. L. Mohlke, L. L. Bonnycastle, et al., "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants," *Science*, vol. 316, no. 5829, pp. 1341–1345, 2007.

[47] R. Saxena, B. F. Voight, V. Lyssenko, et al., "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels," *Science*, vol. 316, no. 5829, pp. 1331–1336, 2007.

[48] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[49] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 3, 2004.

[50] T. R. Golub, D. K. Slonim, P. Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[51] L. J. van't Veer, H. Dai, M. J. van de Vijver, et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[52] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.

[53] L. Xu, A. C. Tan, D. Q. Naiman, D. Geman, and R. L. Winslow, "Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data," *Bioinformatics*, vol. 21, no. 20, pp. 3905–3911, 2005.

[54] Y. Tan, L. Shi, W. Tong, and C. Wang, "Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data," *Nucleic Acids Research*, vol. 33, no. 1, pp. 56–65, 2005.

[55] G. Bloom, I. V. Yang, D. Boulware, et al., "Multi-platform, multi-site, microarray-based human tumor classification," *American Journal of Pathology*, vol. 164, no. 1, pp. 9–16, 2004.

[56] P. Warnat, R. Eils, and B. Brors, "Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes," *BMC Bioinformatics*, vol. 6, article 265, pp. 1–15, 2005.

[57] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Informatics*, vol. 2, pp. 59–78, 2006.

[58] J. R. Stevens and R. W. Doerge, "Combining Affymetrix microarray results," *BMC Bioinformatics*, vol. 6, article 57, pp. 1–19, 2005.

[59] J. Wang, K. A. Do, S. Wen, et al., "Merging microarray data, robust feature selection, and predicting prognosis in prostate cancer," *Cancer Informatics*, vol. 2, pp. 87–97, 2006.

[60] P. Hu, C. M. T. Greenwood, and J. Beyene, "Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models," *BMC Bioinformatics*, vol. 6, articel 128, pp. 1–11, 2005.

[61] R. P. DeConde, S. Hawley, S. Falcon, N. Clegg, B. Knudsen, and R. Etzioni, "Combining results of microarray experiments: a rank aggregation approach," *Statistical Applications in Genetics and Molecular Biology*, vol. 5, no. 1, article 15, 2006.

[62] A. A. Shabalin, H. Tjelmeland, C. Fan, C. M. Perou, and A. B. Nobel, "Merging two gene-expression studies via cross-platform normalization," *Bioinformatics*, vol. 24, no. 9, pp. 1154–1160, 2008.

[63] H. Cooper and L. V. Hedges, *The Handbook of Research Synthesis*, Russell Sage, New York, NY, USA, 1994.

[64] L. V. Hedges and I. Olkin, *Statistical Methods for Meta-Analysis*, Academic Press, Orlando, Fla, USA, 1995.

[65] G. Parmigiani, E. S. Garrett-Mayer, R. Anbazhagan, and E. Gabrielson, "A cross-study comparison of gene expression studies for the molecular classification of lung cancer," *Clinical Cancer Research*, vol. 10, no. 9, pp. 2922–2927, 2004.

[66] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan, "Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer," *Cancer Research*, vol. 62, no. 15, pp. 4427–4433, 2002.

[67] P. Hu, C. M. T. Greenwood, and J. Beyene, "Statistical methods for meta-analysis of microarray data: a comparative study," *Information Systems Frontiers*, vol. 8, no. 1, pp. 9–20, 2006.

[68] X. Yang and X. Sun, "Meta-analysis of several gene lists for distinct types of cancer: a simple way to reveal common prognostic markers," *BMC Bioinformatics*, vol. 8, articel 118, pp. 1–17, 2007.

[69] D. Tritchler, "Modelling study quality in meta-analysis," *Statistics in Medicine*, vol. 18, no. 16, pp. 2135–2145, 1999.

[70] P. Hu, J. Beyene, and C. M. T. Greenwood, "Tests for differential gene expression using weights in oligonucleotide microarray experiments," *BMC Genomics*, vol. 7, article 33, pp. 1–15, 2006.

[71] S. Heber and B. Sick, "Quality assessment of Affymetrix GeneChip data," *OMICS: A Journal of Integrative Biology*, vol. 10, no. 3, pp. 358–368, 2006.

[72] M. Morley, C. M. Molony, T. M. Weber, et al., "Genetic analysis of genome-wide variation in human gene expression," *Nature*, vol. 430, no. 7001, pp. 743–747, 2004.

[73] V. G. Cheung, R. S. Spielman, K. G. Ewens, T. M. Weber, M. Morley, and J. T. Burdick, "Mapping determinants of human gene expression by regional and genome-wide association," *Nature*, vol. 437, no. 7063, pp. 1365–1369, 2005.

[74] B. E. Stranger, M. S. Forrest, A. G. Clark, et al., "Genome-wide associations of gene expression variation in humans," *PLoS Genetics*, vol. 1, no. 6, p. e78, 2005.

[75] P. Hu, H. Lan, W. Xu, J. Beyene, and C. M. T. Greenwood, "Identifying *cis*- and *trans*-acting single-nucleotide polymorphisms controlling lymphocyte gene expression in humans," *BMC Proceedings*, vol. 1, supplement 1, article S7, pp. 1–5, 2007.

[76] E. Parkhomenko, D. Tricheler, and J. Beyene, "Genome-wide sparse canonical correlation of gene expression with genotypes," *BMC Proceedings*, vol. 1, supplement 1, article S119, pp. 1–5, 2007.

[77] J. R. Pollack, T. Sørlie, C. M. Perou, et al., "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 20, pp. 12963–12968, 2002.

[78] A. M. Levin, D. Ghosh, K. R. Cho, and S. L. R. Kardia, "A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors," *Bioinformatics*, vol. 21, no. 12, pp. 2867–2874, 2005.

[79] J. Huang, H.-H. Sheng, T. Shen, et al., "Correlation between genomic DNA copy number alterations and transcriptional expression in hepatitis B virus-associated hepatocellular carcinoma," *FEBS Letters*, vol. 580, no. 15, pp. 3571–3581, 2006.

[80] P. Platzer, M. B. Upender, K. Wilson, et al., "Silence of chromosomal amplifications in colon cancer," *Cancer Research*, vol. 62, no. 4, pp. 1134–1138, 2002.

[81] B. A. Cohen, R. D. Mitra, J. D. Hughes, and G. M. Church, "A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression," *Nature Genetics*, vol. 26, no. 2, pp. 183–186, 2000.

[82] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, vol. 21, supplement 1, pp. i38–i46, 2005.

[83] A. Subramanian, P. Tamayo, V. K. Mootha, et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[84] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *Annals of Applied Statistics*, vol. 1, no. 1, pp. 107–129, 2007.