Published in final edited form as:

Stat Modelling. 2017 ; 17(4-5): 245–289. doi:10.1177/1471082X17698255.

Statistical Contributions to Bioinformatics: Design, Modeling, Structure Learning, and Integration

Jeffrey S. Morris¹ and Veerabhadran Baladandayuthapani¹

¹Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA

Abstract

The advent of high-throughput multi-platform genomics technologies providing whole-genome molecular summaries of biological samples has revolutionized biomedical research. These technologies yield highly structured big data, whose analysis poses significant quantitative challenges. The field of Bioinformatics has emerged to deal with these challenges, and is comprised of many quantitative and biological scientists working together to effectively process these data and extract the treasure trove of information they contain. Statisticians, with their deep understanding of variability and uncertainty quantification, play a key role in these efforts. In this article, we attempt to summarize some of the key contributions of statisticians to bioinformatics, focusing on four areas: (1) experimental design and reproducibility, (2) preprocessing and feature extraction, (3) unified modeling, and (4) structure learning and integration. In each of these areas, we highlight some key contributions and try to elucidate the key statistical principles underlying these methods and approaches. Our goals are to demonstrate major ways in which statisticians have contributed to bioinformatics, encourage statisticians to get involved early in methods development as new technologies emerge, and to stimulate future methodological work based on the statistical principles elucidated in this article and utilizing all available information to uncover new biological insights.

Keywords

Bioinformatics; Epigenetics; Experimental Design; Genomics; Preprocessing; Proteomics; Regularization; Reproducible Research; Statistical Modeling

1 Introduction

Rapid technological advances accompanied by a steep decline in experimental costs have led to a proliferation of genomic data across many scientific disciplines and virtually all disease areas. These include high-throughput technologies that can profile genomes, transcriptomes, proteomes and metabolomes at a comprehensive and detailed resolution unimaginable only a couple of decades ago (Schuster, 2008) – a summary of some of the key technologies is presented in Section 2 and Figure 1. This has led to data generation at an unprecedented

scale in various formats, structure and sizes (Bild et al., 2014; Hamid et al., 2009) – raising a plethora of analytic and computational challenges. The general term *Bioinformatics* refers to a multidisciplinary field involving computational biologists, computer scientists, mathematical modelers, systems biologists, and statisticians exploring different facets of the data ranging from storing, retrieving, organizing and subsequent analysis of biological data. Given the myriad challenges posed by this complex field, bioinformatics is necessarily interdisciplinary in nature, as it is not feasible for any single researcher in themselves to possess all clinical, biological, computational, data management, mathematical modeling, and statistical knowledge and skills necessary to optimally discover and validate the vast scientific knowledge contained in the outputs from these technologies.

Statisticians have a unique perspective and skill set that places them at the center of this process. One of the key attributes that sets statisticians apart from other quantitative scientists is their understanding of variability and uncertainty quantification. These are essential considerations in building reproducible methods for biological discovery and validation, especially for complex, high-dimensional data as encountered in genomics. Statisticians are “data scientists” who understand the profound effect of sampling design decisions on downstream analysis, potential propagation of errors from multi-step processing algorithms, and the potential loss of information from overly reductionistic feature extraction approaches. They are experts in inferential reasoning, which equips them to recognize the importance of multiple testing adjustment to avoid reporting spurious results as discoveries, and to properly design algorithms to search high-dimensional spaces and build predictive models while obtaining accurate measures of their predictive accuracy.

While statisticians have been involved in many aspects of bioinformatics, they have been hesitant to get heavily involved in other aspects. For example, many statisticians are primarily interested in end-stage modeling after all of the data already been collected and preprocessed. Statistical expertise in the experimental design and low-level processing stages are equally if not more important than end-stage modeling, since errors and inefficiencies in these steps propagate into subsequent analyses, and can preclude the possibility of making valid discoveries and scientific conclusions even with the best constructed end-stage modeling strategies. This has resulted in a missed opportunity for the statistical community to play a larger leadership role in bioinformatics that in many cases has been instead assumed by other quantitative scientists and computational biologists, and congruently, a missed opportunity for biologists as well, to more efficiently learn true reproducible biological insights from their data.

With statisticians being a little slow to get involved on the front lines of bioinformatics, many basic and advanced statistical principles have been underutilized in the collection and modeling of bioinformatics data. As a result, we see far too many studies with non-replicated false positive results from confounded experimental designs or improper training-validation procedures, even in high-profile journals. Driven by the computational challenges of high dimensionality and out of convenience, many commonly used standard analysis approaches are reductionistic (not modeling the entire data set), *ad hoc* and algorithmic (not model-based), stepwise and piecemeal (not integrating information together in a statistically efficient way or propagating uncertainty through to the final analysis). Failure to use all of

the information in the data increases the risk of missed discoveries. Greater involvement by the statistical community can help improve the efficiency and reproducibility of genomics research.

In spite of missed opportunities, there have been substantial efforts and success stories where (sometimes advanced) statistical tools have been developed for these data, leading to improved results and deep scientific contributions. The goals of this article are to highlight how statistical modeling has and can make a difference in bioinformatics, elucidate the key underlying statistical principles that are relevant in many areas of bioinformatics, and stimulate future methodological development. While drawing some general conclusions, we organize the core of this paper around four key areas:

1. Experimental design and reproducible research (Section 3)
2. Improved preprocessing and feature extraction (Section 4)
3. Flexible and unified modeling (Section 5)
4. Structure learning and integration (Section 6)

In this article, we do not attempt to exhaustively summarize the work that has been done, but instead attempt to illustrate contributions and highlight the motivating statistical principles. For each area, we will present case examples, including some high profile research that has significantly impacted bioinformatics practice as well as other works (including some of our own) that illustrate some of the key statistical principles even if not as impactful. Through these examples we will extract and highlight some of the key underlying statistical principles, including randomization and blocking, denoising, borrowing strength across correlated measurements, and unified modeling. Our hope is that the elucidation of these principles will help guide and stimulate future methodological development and increase the role and impact of statistics in bioinformatics.

2 Background

As alluded to above, the advent of high-throughput molecular technologies has revolutionized biomedical science. In this section, we overview some basic data structures generated by molecular platforms at different resolution levels, including mRNA-based (transcriptomics), DNA-based (genomics), protein-based (proteomics), as well as epigenetic factors (e.g. methylation). The key underlying principle is that the biological behavior of cells, normal or diseased, is regulated by molecular processes, and different aspects of these processes are measured by these various platforms.

The basic information flow across the various resolution levels starts with a gene encoded by DNA to messenger RNA (mRNA) by a biological process called *transcription* and mRNA to protein by a process called *translation*. This basic process can be regulated and altered through epigenetic processes such as DNA methylation that help regulate transcription, post translational modification of histone proteins within the chromatin structures encasing the DNA, or by micro RNAs (miRNAs) that degrade targeted mRNA. It is through these molecular activities that biological processes are regulated and phenotypes within organisms

are determined. A simplified model illustrating these interrelationships is given by Mallick et al. (2009).

$$\text{DNA} \rightarrow \text{mRNA} \rightarrow \text{Protein} \rightarrow \text{Cell phenotype} \rightarrow \text{Organism phenotype}$$

More complicated relationships and feedback loops are being discovered and work off of this fundamental information flow. A schematic of the various data types and some of their interrelationships is provided in Figure 1.

To provide a backdrop for our discussion, we briefly overview each of these molecular resolution levels and explain some of the specific data structures generated by the corresponding assays. To make the article accessible to quantitative/statistical readers we eschew some of the biological/technical details and refer the readers to specific references in each section.

2.1 Transcriptomics

Early work in measuring gene mRNA expression data were based on a “one-gene-at-time” process using hybridization based methods (Gillespie and Spiegelman, 1965) such as Northern Blots (Alwine et al., 1977) and Reverse transcription polymerase chain reaction (RT-PCR) experiments. Broadly, the purpose of these low-throughput experiments was to measure the size and abundance of the RNA transcribed for an individual gene using cellular RNA extraction procedures applied to multiple cells from a organism or sample. These experiments were typically time-consuming and involved selection of individual genes to assay expression and were mostly used for hypothesis driven endeavors.

The advent of microarray-based technologies in the mid-1990’s then automated these techniques to simultaneously measure expressions of thousands of genes in parallel. This shifted gene expression analyses from mostly hypothesis driven endeavors to hypothesis generating ones that involve an unbiased exploration of the expression patterns of the entire transcriptome.

The major types of arrays can be predominantly classified into three main categories (see Seidel (2008) for detailed review). The first reported works in microarrays involved **spotted microarrays** developed at Stanford University (Schena et al., 1995; Shalon et al., 1996; Schena et al., 1998). Broadly, the process involves printing libraries of PCR products or long oligonucleotide sequences from a set of genes onto glass slides via robotics and then estimating the gene expression intensities through fluorescent tags (see Brown and Botstein (1999) for a detailed review). Other institutions developed laboratories for printing their own spotted microarrays, which had variable data quality given the challenge of reproducible manufacture of the arrays. **Affymetrix** was the among the first company to standardize the production of microarrays, becoming the most established and widely-used commercial platform for measure high-throughput gene expression data. Their arrays consist of 25-mer oligo-nucleotides synthesized on a glass chip (Pease et al., 1994). As opposed to the single sequence of probes used in spotted microarrays, Affymetrix uses a set of probes to measure and summarize expression of the genes. Subsequently, other companies, including Illumina,

Agilent and Nimblegen, have produced microarrays involving *in situ* synthesis, with each using a different type and length of oligonucleotide as well as photo-chemical process for measurement of gene expression (Blanchard et al., 1996). As described in the next section, the development of more efficient and cost-effective sequencing technologies has led to the use of next generation sequencing (NGS) technologies applied to RNA, RNAseq, as the preferred mode of measuring gene expression.

While each technology has its own characteristics and caveats (see Section 4.1), the basic read outs contain expression level estimates for thousands on genes on a per-sample basis. This has been used for discovery of the relative fold change in disease versus normal tissues (Alizadeh et al., 2000) and among different disease tissue types (Ramaswamy et al., 2001). Moreover, these technologies have been used to discover molecular signatures that can differentiate subtypes within a given disease that are molecularly distinct (Bhattacharjee et al., 2001; DeRisi et al., 1997; Eisen et al., 1998; Guinney et al., 2015). Clinical applications include but are not limited to development of diagnostic and prognostic indicators and signatures (Cardoso et al., 2008; Bueno-de Mesquita et al., 2007; Bonato et al., 2011).

2.2 Genomics

Taking a step back, DNA-based assays measure genomic events at the DNA level before transcription. Relevant DNA alterations include natural variability in germline genotype, or the DNA sequence, across individuals that sometimes affect biological function and disease risk, as well as germline or somatic genomic aberrations including various types of mutations, including substitutions, insertions, deletions and translocations, as well as broader changes in the genome including loss of entire chromosomes or parts of a chromosome or loss of heterozygosity (LOH), involving the loss of one of two distinct alleles originally possessed by the cell.

Diploid organisms such as humans have two copies of each autosome (i.e. non-sex chromosomes), but many diseases are associated with aberration in the number of DNA copies in a cell, especially cancer (Pinkel and Albertson, 2005). Most diseases acquire DNA copy number changes manifesting as entire chromosomal changes, segment-wise changes in the chromosome, or modification of the DNA folding structure. Such cytogenetic modifications during the life of the patient can result in disease initiation and progression by mechanisms wherein disease-suppression genes are lost or silenced, or promoter genes that encourage disease progression are amplified. The detection of these regions of aberration has the potential to impact the basic knowledge and treatment of many types of diseases and can play a role in the discovery and development of molecular-based personalized therapies.

In early years, cytogeneticists were limited to visually examining whole genomes with a microscope, a technique known as *karyotyping* or *chromosome analysis*. In mid-70's and 80's the development and application of molecular diagnostic methods such as Southern blots, polymerase chain reaction (PCR) and fluorescence in situ hybridization (FISH) allowed clinical researchers to make many important advances in genetics, including clinical cytogenetics. However these techniques have several limitations. First, they are very time-consuming and labor-intensive, and only a limited number and regions of the chromosome can be tested simultaneously. Further, because the probes are targeted to specific

chromosome regions, the analysis requires prior knowledge of an abnormality and was of limited use for screening complex karyotypes. More recently scientists have developed techniques that integrate aspects of both traditional and molecular cytogenetic techniques called *chromosomal micorarrays* (Vissers et al., 2010). These high-throughput high-resolution microarrays have allowed researchers to diagnose numerous subtle genome-wide chromosomal abnormalities that were previously undetectable and find many cytogenetic abnormalities in part or all of a single gene. Such information is useful for biologists to detect new genetic disorders and also provide better understanding of the pathogenetic mechanisms of many chromosomal aberrations.

Broadly, there are two types of chromosomal microarrays: array-based Comparative genomic hybridization (aCGH arrays) and single nucleotide polymorphism microarrays (SNP arrays). CGH-based methods were developed to survey DNA copy number variations across a whole genome in a single experiment (Kallioniemi et al., 1992) With CGH, differentially labeled test (e.g., tumor) and reference (e.g., normal individual) genomic DNAs are co-hybridized to normal metaphase chromosomes, and fluorescence ratios along the length of chromosomes provide a cytogenetic representation of the relative DNA copy number variation. Chromosomal CGH resolution is limited to 10–20 Mb, hence any aberration smaller than that will not be detected. Array-based comparative genomic hybridization (aCGH) is a subsequent modification of CGH that provided greater resolution by using microarrays of DNA fragments rather than metaphase chromosomes (Pinkel et al., 1998; Snijders et al., 2001). These arrays can be generated with different types of DNA preparations. One method uses bacterial artificial chromosomes (BACs), each of which consists of a 100- to 200-kilobase DNA segment. Other arrays are based on complimentary DNA (cDNA, Pollack et al. (1999)) or oligonucleotide fragments (Lucito et al., 2000). As in CGH analysis, the resultant map of gains and losses is obtained by calculating fluorescence ratios measured via image analysis tools.

SNP arrays are one of the most common types of high-resolution chromosomal microarrays (Mei et al., 2000). SNPs, or single nucleotide polymorphisms, are single nucleotides in the genome in which variability across individuals or across paired chromosomes has been observed. Researchers have already identified more than 50 million SNPs in the human genome. SNP arrays take advantage of hybridization of strands of DNA derived from samples, each with hundreds of thousands of probes representing unique nucleotide sequences. As with aCGH, SNP-based microarrays quantitatively determine relative copy number for a region within a single genome. Platform-specific specialized software packages are used to align the SNPs to chromosomal locations, generating genome-wide DNA profiles of copy number alterations and allelic frequencies that can then be interrogated to answer various scientific and clinical questions.

Note that unlike aCGH arrays, SNP arrays have the advantage of detecting both copy number alterations as well as LOH events given the allelic fractions, typically referred to as the B-allele frequencies (Beroukheim et al., 2006). They also provide genotypic information for the SNPs, which when considered across multiple SNPs can be used to study haplotypes. SNP array analysis of germline samples have been extensively used in genome-wide association studies (GWAS) to find genetic markers associated with various disease of

interest. We refer the reader to Yau and Holmes (2009) for nice review of the data elements obtained via SNP arrays.

The initial human genome project involved a complete sequencing of a human genome, which took 13 years (1990–2003) and cost roughly \$3 billion. Over the past decade, great improvements have been made in the hardware and software undergirding sequencing, leading to next generation sequencing (NGS) that can now be used to sequence an entire human genome in less than a day for a cost of about \$1000. This sequencing data obtained by applying NGS to DNA, DNAseq, can be used to completely characterize genotypes in GWAS studies, and to characterize genetic mutations for diseased tissue such as cancer tumors. Many types of mutations can be characterized, including point mutations, insertions, deletions, and translocations. DNAseq can also be used to estimate copy number variation and LOH throughout the genome. Cost and time of sequencing is largely determined by depth of sequencing, e.g. with 30× depth indicating that we expect to get at least 30 counts of each genomic location. When focus is on common mutational variants and copy number determination, low depth sequencing (8x–10x) may be sufficient, but much higher depth is required if rare variants are to be detected. Also, at times targeted sequencing is done to focus on specific parts of the genome, e.g. whole exome sequencing for which the gene coding regions only are sequenced.

2.3 Proteomics

Proteomic technologies allow direct quantification of protein expression as well as post-translational events. Although much more difficult to study than DNA or RNA because their abundance levels span many orders of magnitude, it is important to study proteins as these play a functional role in cellular processes and numerous studies have found that mRNA expression and protein abundance often correlate poorly with each other. Here, we will briefly overview several important proteomic technologies that involve estimating absolute or relative abundance levels, including low to moderate-throughput assays that can be used to study small numbers of pre-specified proteins and high-throughput methods that can survey a larger slice of the proteome.

Low to moderate-throughput proteomic assays—Traditional low-throughput protein assays include immunohistochemistry (IHC), Western blotting and enzyme-linked immunosorbent assay (ELISA). Although IHC is a very powerful technique for the detection of protein expression and location, it is critically limited in statistical analyses by its non- to semi-quantitative nature. Western blotting can also provide important information, but due to its requirement for relatively large amounts of protein, it is difficult to use when comprehensively assessing large-scale proteomic investigations, and also is semi-quantitative in nature. The ELISA method provides quantitative analysis, but is similarly limited by requirements of relatively high amounts of specimen and by the high cost of analyzing large pools of specimens.

To overcome these limitations, Reverse-phase protein arrays (RPPA) have been developed to provide quantitative, high-throughput, time- and cost-efficient analysis of small to moderate number of proteins (dozens to hundreds) using small amounts of biological material (Tibes

et al., 2006). In RPPA analyses, proteins are isolated from the biological specimens such as cell lines, tumors, or serum using standard laboratory-based methods. The protein concentrations are then determined for the samples and subsequently, serial 2-fold dilutions prepared from each sample are then arrayed on a glass slide. Each slide is then probed with an antibody that recognizes a specific protein epitope, that reflects the activation status of the protein. A visible signal is then generated through the use of a signal amplification system and staining. The signal reflects the relative amount of that epitope in each spot on the slide. The arrays are then scanned and the resulting images are analyzed with an imaging software (MicroVigene, VigeneTech Inc., Carlisle, MA) that can be used to quantify protein abundance for each protein for each sample. We refer the reader to Paweletz et al. (2001) and Hennessy et al. (2010) for more biological and technical details and Baladandayuthapani et al. (2014) for quantitative details concerning RPPAs.

High throughput proteomic assays—While RPPAs are useful for studying pre-specified panels of proteins, at times researchers would like to assess proteomic content and abundance on a larger and unbiased scale using high-throughput technologies. 2D gel electrophoresis was developed in the 1970's (O'Farrell, 1975), and has served as the primary workhorse for high-throughput expression proteomics. 2DGE physically separates the proteomic content of a biological sample on a polyacrimide gel based on isoelectric point (pH) and molecular mass, which is then scanned. The resulting gel image is characterized by hundreds or thousands of spots each corresponding to proteins present in the sample that are analyzed to assess protein differences across samples. Because the spots on the gel contain actual physical proteins, the proteomic identity of a spot can be determined by cutting it out of the gel and further analyzing it using protein identification techniques like tandem mass spectrometry (see below). A variant of 2DGE that can potentially lead to more accurate relative abundance measurements is 2D difference gel electrophoresis (DIGE, Karp and Lilley (2005)), which involves labeling two samples with two different dyes, loading them onto the same gel, and then scanning the gel twice using different lasers that differentially pick up on the two dyes. This can be used in paired designs to find proteins with differential abundance between two conditions, or in more general designs a common reference material can be used on the second channel as an internal reference factor.

Recently, 2DGE has fallen out of fashion for high-throughput proteomics, at least partly because of the lack of automatic, efficient, and effective methods to analyze the gel images. Mass spectrometry (MS) approaches have gained prominence in its place. Mass spectrometry methods survey the proteomic content of a biological sample by measuring the mass-per-unit charge (m/z) ratio of charged particles. Various technologies exist, which vary in terms of the approach used to generate the ions (e.g. MALDI, matrix-assisted laser desorption and ionization, and ESI, electrospray ionization) and to separate the proteins based on their molecular mass (e.g. TOF, time of flight, QIT, quadruple ion trap, FT-ICR, Fourier-transform ion cyclotron resonance). In each case, the separated ions are detected and assembled into a mass spectrum, a spiky function that measures abundance of particles over a series of time points, which can subsequently be mapped to m/z values. While commonly used for protein identification, relative protein abundances can also be assessed and compared between groups via quantitative analysis of the spectra.

Given the large number of proteins present in a sample at varying abundance levels spanning many orders of magnitudes, it is not possible to survey all proteins in a single spectrum. Liquid chromatography (LC) is combined with mass spectrometry to survey a larger slice of the proteome. In LC-MS, proteins are digested and separated in an LC column based on the gradient of some chosen factor (e.g. hydrophobicity). Over a series of elution times, the set of separated proteins are then fed into a MS analyzer to produce a spectrum. This technique effectively separates the proteins based on two factors (e.g. hydrophobicity and m/z), and can be visualized as “image data” with “spots” consisting of m/z peaks over a series of elution times corresponding to particular proteins. Commonly, a second MS step is done to produce protein identifications and peptide counts for a subset of peaks at each elution time, in which case the approach is called LC-MS/MS. While taking a long time to run, these techniques show promise for broad proteomic characterization of a biological sample.

2.4 Epigenetics

Although the central dogma of genetics is that DNA is transcribed into mRNA which is then translated into proteins, this process is regulated and can be altered by many other molecular processes that affect gene expression but are not directly related to the genetic code. The study of these processes is called *epigenetics*, and includes processes such as methylation, histone modification, transcription factor binding, and micro RNA (miRNA) expression.

One of the major and most studied epigenetic factors is methylation, whereby a methyl group is added to DNA at a CpG site in which a cytosine is connected to a guanine by a phosphodiester bond. This methylation can alter gene expression, for example by repressing transcription especially when located near the promoter region of the gene, but methylation at other locations can also affect gene expression in various ways. Methylation can be modified by environmental factors, and modifications are usually inherited through mitosis and sometimes even meiosis, so can permanently alter gene expression and is an important component of many diseases, including cancer.

One common approach to measure methylation is to use sodium bisulfite conversion, in which sodium bisulfite added to DNA fragments that converts unmethylated cytosine into uracil, allowing the estimation of a *beta value* measuring the percent methylation at a given CpG site. This technique can be used on individual CpGs, or has been used to generate methylation arrays capable of surveying the entire genome. In 2009, Illumina introduced a 27k array (Bibikova et al., 2009) that measured methylation at 27,578 CpG sites from 14,495 genes, with roughly two CpG per genes. This array focused on promoter regions, including CpG islands, genomic regions containing a high frequency of CpG sites (Bird et al., 1987). Early methylation research focused on CpG islands, which were thought to be the most important regulatory regions. However, a team led by biostatistician Rafael Irizarry (Irizarry et al., 2009) (1317 citations, Google Scholar) discovered by empirical analyses of a broader array of CpG sites in the genome that most methylation alterations that separate different types of tissues and characterize differences between normal tissue and cancer do not occur on these CpG islands, but in sequences up to 2kb distant from CpG islands, which they term *CpG shores*. This discovery fueled further development of tools to more broadly survey methylation across the genome, first with Illumina developing a 450k array (Bibikova

et al., 2011) that included 487,557 CpGs from many different locations in the genome, and whole genome bisulfite sequencing (WGBS) (Lister et al., 2009) which uses bisulfite conversion and NGS to obtain beta values for every CpG in the genome.

Some other epigenetic factors include histone modification, transcription factor binding, and miRNA expression. DNA is contained within chromatin structures in which DNA wraps around histone proteins forming nucleosomes. These histones can be modified by processes such as acetylation, phosphorylation, methylation, and deimination, that can affect DNA expression (Bannister and Kouzarides, 2011). While known to be important, the full functional characterization of these modifications is still being discovered. Histone modification status can be measured genome-wide using chromatin immunoprecipitation (ChIP) (Collas, 2010). Transcription is typically initiated by the binding of a protein known as a transcription factor to a binding site close to the 5' end of the gene, and the study of which transcription factors affect which genes has important functional implications. *ChIP-seq* is a modern tool that combines chromatin immunoprecipitation with sequencing to find binding sites for transcription factors. MicroRNAs (miRNA) are short, single-stranded fragments of non-coding RNA that are involved in gene expression regulation. Typically, a miRNA functions by binding to a target sequence and degrading a set of target mRNA, inhibiting translation. miRNA can be measured like other mRNA, including real-time quantitative PCR, microarrays, and RNAseq.

2.5 Data structures, characteristics and modeling challenges

The data structures emanating from these high-throughput technologies have various explicit and implicit *structured dependencies*, some caused by underlying biological factors and others technically induced by experimental design – which can have profound implications in downstream modeling. We attempt to characterize some major modes of these dependencies. Transcriptomic and proteomic data typically generate large scale multivariate data with large number of variables (genes/proteins), typically much higher orders than the sample size – the large p , small n situation, which is a common thread in nearly all of the technologies described above. Copy number and methylation typically generate profiles, indexed by genomic location, hence inherently exhibit serial or spatial correlations which can be both short and long range. In addition, the data structures can be on vastly different scales, continuous (e.g. protein/gene expression), discrete (e.g. copy number states), count data (e.g. RNA sequencing) and measurements on bounded intervals (e.g. methylation status $\in (0, 1)$).

Furthermore, the underlying biological principles induce a natural higher order organization to these variables – such as grouping based on common biological functions of the genes/proteins and complex regulatory signaling and mechanistic interactions between them. These can induce dependencies in the data across genes/proteins/genomic locations. This raises modeling and inferential challenges that requires the appropriate level of sophistication, ranging from pre-processing to downstream modeling, that we detail in the following sections.

3 Experimental Design, Reproducible Research, and Forensic Statistics

The importance of statistical input into experimental design has been known for some time. In the early 1930's, R.A. Fisher famously said, "*To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.*" This is even more true for modern experiments involving highly sensitive technologies generating complex high dimensional data. The involvement of statisticians in the design phase can make fundamental contributions to science by helping ensure reproducible research.

Reproducibility and replicability of results are key elements of scientific research. The field of multi-platform genomics has been plagued with lack of reproducibility (Ioannidis et al., 2009; Begley and Ellis, 2012), much of which can be explained by two primary factors: (1) sensitivity of the technologies to varying experimental conditions or sample processing and (2) the inherent complexity of the data leading to poorly documented and sometimes flawed analytical workflows and a neglect to do due diligence in exploratory data analysis. In this section, we will summarize two case studies in which re-analyses of data from seminal publication by statisticians revealed erroneous results, and effectively served as the type of post-mortem analysis mentioned by Fisher that could be called *forensic statistics*. We will describe how these efforts illustrate the importance of applying fundamental statistical principles of experimental design and exploratory data analysis (Tukey, 1977) to multi-platform genomics, and have provided an impetus to major scientific journals and federal agencies to establish policies to ensure greater reproducibility of research, especially for high-throughput multi-platform genomics data.

3.1 OvaCheckTM: Proteomic Blood Test for Ovarian Cancer

In 2002, a paper in the *The Lancet* (Petricoin et al., 2002) reported a blood test based on proteomics mass-spectrometry could be used to detect ovarian cancer with near 100% sensitivity and specificity. If true, this could revolutionize the management of ovarian cancer, as this lethal disease is typically detected in later stages when treatments are generally ineffective, and at the time there were no reliable screening techniques for early detection of this disease. This generated a great deal of interest in the medical community, and the researchers developed a commercial blood test, OvaCheckTM, that was to become available to women nationwide in early 2004.

This also generated a great deal of interest at M.D. Anderson Cancer Center, as many cancer researchers wanted to try this approach for early detection of other cancers and came to the Biostatistics department for help planning these studies. With no previous experience with mass spectrometry data, Kevin Coombes, Keith Baggerly, and Jeffrey Morris set out to understand these data so we could assist our collaborators on doing similar studies. Fortunately, the authors of the seminal *Lancet* paper made their data publicly available, so we downloaded it with the intention of familiarizing ourselves with the data and figuring out how to analyze it. However, instead we ended up uncovering some serious questions about the data and the veracity of the published results.

In the initial paper, they had blood serum from 100 healthy subjects, 100 ovarian cancer patients, and 16 patients with benign ovarian cysts. They split off 50 healthy subjects and 50 ovarian cancer subjects and trained a classifier using proteomic features from the mass spectrometry, and then for validation, they applied to the additional samples and correctly classified all 50 cancers as cancers, 47/50 of the normals as normal, and remarkably 16/16 of the benign cysts as neither cancer nor normal. These data were obtained from running the samples on a Ciphergen H4 ProteinChip array, and then the samples were rerun on a Ciphergen WCX2 ProteinChip array, which binds a different subset of the proteome, and both of these data sets were posted on the web. One of the first things we did after downloading these data was plot a heatmap of each of them, see Figure 2. Based on this heatmap, it became evident that the benign cysts were very different from both cancers and normals, and that the benign cyst mass spectra from the first data set looked much like the WCX2 array data from the second study. Thus, it appeared that the benign cyst samples in Petricoin et al. (2002) were in fact run on a different Ciphergen chip, and their correct classification as neither cancer nor normal was driven by this technical artifact, not biology. This was disturbing, as the article stated that “positives and controls were run concurrently, intermingled on the same chip and multiple chips,” and surprisingly, no one had caught this error before our analysis. This demonstrates the importance of exploratory data analysis (Tukey, 1977).

Data from another follow-up study by this group was also posted on the web. This data set contained spectra from running the Ciphergen WCX2 ProteinChip array on 91 healthy subjects and 162 subjects with ovarian cancer, independent from the original data set. Near-perfect classification was also reported for these data, although the mass spectrometry protein features reported were different from those reported in Petricoin et al. (2002). We performed numerous analyses that demonstrated that both data sets seemed to discriminate cancer from normal when cross validation was used for evaluation, but the classifiers derived from one data set did not discriminate for the other data set (Baggerly et al., 2004b, 2005a,b). A simulation study (Baggerly et al., 2005b) revealed that the spectra in the second data set had pervasive differences between cancer and normal, which should not be the case since we expect that biological proteomic differences should be characterized by a limited number of specific peaks, not the entire spectrum. In spite of the authors originally asserting that the cases and controls were co-mingled on the arrays for these data (Petricoin et al. (2004), appeared in print as commentary to Sorace and Zhan (2003)), they later acknowledged that in fact “case and control samples were run in separate batches on separate days” (Liotta et al., 2005), which would explain such pervasive differences. We also confirmed such confounding of run order and case/control status in another follow up study by this group (Baggerly et al., 2004a). Such confounding can hard-code spurious signals between cases and controls, and is a major source of irreproducible results in science.

A month after publication of Baggerly et al. (2004b), the FDA sent a letter to the company to hold off on marketing OvacheckTM, and six months later they notified them of the need to conduct further validation studies, studies which never were able to reproduce the initial spectacular results. The involvement of statisticians in this type of forensic statistical analysis revealed that the initial remarkable results were spurious artifacts of severe design flaws in these studies and prevented an ineffective diagnostic from going to market.

3.2 Duke Scandal: Predicting Response to Cancer Therapy

Potti et al. (2006) introduced a new strategy for using information from microarrays run on the NCI60 cell lines to build predictive signatures to determine which patients are likely to respond to which cancer therapy. Their strategy was to use drug panels to select the 10 most sensitive and 10 most resistant cell lines for a given therapy, train a classifier using the publicly available microarray data for these cell lines, and then apply this model to patient microarray data to obtain a prediction of response. They reported remarkable success of this approach for a number of common chemotherapy agents, including docetaxel, doxorubicin, paclitaxel, 5-FU, cyclophosphamide, etoposide, and topotecan, and some combination therapies.

This work generated a lot of excitement in the medical community, and was named one of the “Top 6 Genetic Stories of 2006” (Discover, 2007). These authors successfully repeated this strategy in other settings, including cisplatin and pemetrexed (Hsu et al., 2007), several combination therapies in breast cancer (Bonnet et al., 2007), and temozolomide (Augustin et al., 2009), with over 15 papers published on this approach over a three year period and clinical trials commenced to prospectively test them. The researchers, university, and other collaborators started up a company that could market this as a potential clinical decision making tool. Despite the excitement generated by these apparent successes, no other groups were able to get this strategy to work, in spite of its sole reliance on publicly available microarray data and cell lines that anyone could have obtained. Perhaps this should have been a warning sign that these results were not all they appeared to be.

Again in the backdrop of this hype, M.D. Anderson investigators wanted to utilize this strategy and so reached out to faculty in the Department of Bioinformatics and Computational Biology for assistance in designing these studies. Keith Baggerly and Kevin Coombes set out to understand these studies so they could replicate them in other settings. However, poor documentation turned this effort into an intensive post-facto reconstruction, and another case of forensic statistics.

Their analysis, which ultimately took 1000s of hours, uncovered a slew of data handling and processing errors including scrambling of gene and group (sensitive/resistant) labels, design confounding, inclusion of genes of unknown origin, and figure duplication. One common error was the swapping of sensitive/resistant labels in the training data, leading to signatures that would in fact propose treatments that patients are least likely to benefit from. There also were numerous irregularities in labeling of the test set, with some mislabeled and others used multiple times in the same analysis, which led to inaccurate reports of the methods’ performance. One key study included four key signature genes whose origins are unknown, as they were clearly not produced by the software and two were not even on the microarrays used for the NCI60 data set. One figure in a later publication duplicated the figure from an earlier publication with a completely different treatment. In one case, the authors asserted that they were blinded to the clinical response in the test data (*The Cancer Letter*, October 2, 2009), but collaborators who sent the data disputed this assertion and instead claimed that the study was not blinded and further they could not replicate the authors’ findings themselves (*The Cancer Letter*, October 23, 2009). Baggerly and Coombes had numerous interactions with the senior authors, Potti and Nevins, but they were unable or unwilling to

address the most serious of these issues and eventually stopped responding. Baggerly and Coombes (2010) reported these irregularities in an *Annals of Applied Statistics* paper, showed that their attempts to follow the reported procedures resulted in predictive results no better than random chance, and strongly urged that clinical trials that had commenced to test these signatures be suspended until these irregularities are rectified.

Shortly after publication of this paper, three clinical trials based on these studies were suspended and a fourth terminated (*The Cancer Letter*, October 9 and 23, 2009), and Duke University convened a panel of outside experts to investigate this research and the original results. Surprisingly, this panel determined that the concerns were unfounded and the trials were restarted. Apparently, the panel was never shown the full list of irregularities raised by Baggerly and Coombes, but only a hand-picked subset that were more easily addressable. It appeared that these concerns were falling on deaf ears and this work was going to be allowed to continue, until it was discovered that one of the senior authors on the work, Anil Potti, had falsified information on his vitae. This provided impetus to review the work more carefully, and Baggerly and Coombes' concerns were verified, and led to strong suggestions of data manipulation and research misconduct. All of this resulted in the ultimate shutdown of all of these trials, a large court settlement from Duke University to families of patients who participated in these trials, and the tarnishing of the reputation of many researchers associated with them. This saga received a great deal of coverage in the national and international media, with 60 Minutes doing a story on it (<https://archive.org/details/KPIX2012021303000060Minutes>) in 2012 that featured Baggerly and Coombes. Once again, a group of statisticians performing forensic statistics uncovered flaws in a high profile study, in this case exposing serious data provenance and integrity issues, and sparing patients exposure to ineffective clinical devices based on the spurious results.

3.3 Lessons Learned: Statistics and Reproducibility

These case studies highlight the perils of research based on complex, high-dimensional data, and the crucial role good statistical practice plays in reproducible research. The first case study highlights the importance of exploratory data analysis and sound experimental design. Many researchers neglect to examine basic graphical summaries of their data prior to analysis, presumably because of their high dimensionality and complexity. The false positive results reported in the ovarian cancer study (Petricoin et al., 2002) would have never seen the light of day had the investigators simply looked at heat maps of the raw spectra plotted above that took less than five minutes to produce. Tukey (1977) highlights the importance of good exploratory data analysis in statistics, and this is not any less true in the world of big data. The first case study demonstrated the hazards of neglecting this aspect of statistical analysis.

Second, it is tempting to use convenience designs in running large scale genomics and proteomics studies since their assays feature complex, multi-step laboratory procedures. However, the great biological sensitivity that makes these assays so desirable for research can also make them highly sensitive to variability in experimental conditions or sample handling. Thus, it is crucial to think carefully about how the experiment is run, with special care taken to avoid confounding the factors of interest with technical factors. This

confounding of case control status with run order was the fatal flaw leading to the false positive results in the first case study. This means that sample handling needs to be consistent across cases and controls, and one must be careful to not confound case and control status with run order. Randomized block designs should be used whenever possible, and diagnostics such as cluster and principal component analyses should be used to examine potential effects of batch and other technical factors. These principles are fundamental to the field of statistics, yet in many cases underutilized in genomics, indicating there is a greater need for statistical involvement in the research process to prevent the need for forensic statistical analyses.

The second case study highlighted the importance of careful documentation of analytical steps in sufficient detail so that the results in the publication can be reproduced given the raw data. This documentation should include any preprocessing steps, gene selection, model training, and model validation procedures. The lack of such documentation made it necessary for Baggerly and Coombes to spend 1000s of hours to reconstruct what was done. In most cases, such efforts are not feasible, and thus there could be many other pivotal studies with spurious or otherwise erroneous results that are allowed to stand, and countless resources spent trying to replicate or build on these results, and in some cases patient treatment decisions being based upon them. Many journals have introduced guidelines requiring a greater level of data sharing and documentation of design and analytical details in the supplementary materials. These high profile re-analyses by statisticians have contributed to a greater level of awareness of these crucial issues, and statisticians have been taking leadership roles in helping funding agencies and journals to construct policies that contribute to greater transparency and reproducibility in research (Peng, 2009; Stodden et al., 2013; Collins and Tabak, 2014; Fuentes, 2016; Hofner et al., 2016)

3.4 Proper Validation and Multiple Testing Adjustment

The second case study also highlighted the importance of proper validation of predictive models. In that case study, there were fundamental problems involving data provenance and blinding irregularities, but in many other cases models built from high-dimensional genomics data are not properly validated for other reasons. Given a large number of potential predictors, it is very easy in high dimensional settings to arrive at a model with excellent or even perfect predictive accuracy for the training data. Thus, it is especially important to properly validate these predictive models to ensure their predictive performance is not biased as a result of overfitting the training data. This can be done by splitting a single data set into training and validation data through cross validation. However, as seen in the first case study, this cannot overcome technical artifacts that might be hard-wired into the entire data set, and so it is far preferable to validate using a second independent data set whenever available. In either case, it is important to ensure all gene selection and modeling decisions are made using the training data alone. A common erroneous practice is to use combined training-validation data set for some modeling decisions such as gene selection, and then only truly validate the parameter estimation step. Since it is the gene selection and not the parameter estimation that typically introduces the most variability into the modeling, this practice can lead to strongly biased predictive accuracy assessments.

A related aspect of reproducible research for high-dimensional genomics and proteomics data in which statisticians have played a crucial role is multiple testing adjustment. In the early days of microarrays, researchers would flag genes as “differentially expressed” based on non-statistical measures like fold-change that ignore within-group variability, or after applying independent statistical tests to 1000s or more of genes while using a standard 0.05-level significance level, leading to high false discovery rates. Early work by statisticians emphasized the importance of using statistical tests and to adjust for multiple testing. For example, Dudoit et al. (2002) presented various methods that strongly controlled family-wise error rate (FWER), including Bonferroni and also step-down strategies that accounted for intergene correlation and thus were less conservative. These initial solutions, however, were not so well received by the biological community because of the low power resulting from their strict experimentwise error rate criteria. Against this backdrop, researchers turned to use of false discovery rate (FDR), a concept introduced by Benjamini and Hochberg (1995) that controls the proportion of false discoveries rather than the probability of at least one false discovery. Researchers broadly deemed this a more appropriate statistical criterion for discovery involving high-dimensional genomics and proteomics data. The statistical community has developed a whole set of tools for FDR analyses, including seminal frequentist methods (Storey, 2002, 2003; Efron, 2004) as well as some Bayesian methods (Newton et al., 2004; Muller et al., 2004; Morris et al., 2008b) for multiple testing adjustment. Statisticians have been able to successfully communicate the necessity of accounting for multiple testing to the broader genomics and proteomics communities, which has to some degree helped mitigate the publishing of false positive discoveries.

4 Improved Preprocessing and Feature Extraction

Several processing steps need to be applied to the raw data generated by the genomics platforms described in Section 2 before they are ready for downstream statistical analysis. While the particulars are technology-specific, there are several general considerations that apply to nearly all technologies. These include additive correction for background signal, multiplicative normalization to account for factors such as variability in amount of biological material loaded or contrast on an optical scanner, and registration of the data so cognate features are aligned across replicates. Batch correction can be a major consideration when samples are collected or processed over a period of time or at different locations. If these steps are not performed well, then technical artifacts can creep into the data and make it difficult to extract biological information from them no matter how well thought out the downstream statistical analysis plan. While many statisticians may consider these preprocessing issues to be low-level “data cleaning” and not fundamental research problems, they are quantitatively challenging, and of primary importance to molecular biology. In a number of cases statisticians have devised preprocessing tools that have made strong contributions to the field. We highlight some of these contributions in this section.

The data generated by many of the high throughput assays described above are complex and high dimensional, and in many cases can be characterized as highly structured functional and image data. A book by Ramsay and Silverman (1997) popularized the concept of *functional data analysis*, which involves treating functional objects as single entities rather than just a collection of discrete data points. These functional objects can be simple smooth

curves on one-dimensional Euclidean domains, or can be more complex objects with local features, and potentially defined on higher dimensional domains or non-Euclidean manifolds, as many types of modern data including various types of high-throughput genomics. One strategy for modeling complex functional data is to simplify the data using a *feature extraction approach*, which is a two-step approach whereby first, statistical summaries believed to capture the key biological information in the data are computed, and then second, these summaries are modeled using standard statistical analysis tools.

This is the predominant analytical strategy for nearly all of the high throughput assays described in Section 2. Examples include aggregation of information across multiple probes or genomic locations to produce gene expression summaries, performing peak or spot detection in proteomics to obtain counts or relative protein abundance measurements, or segmenting regions of the genome believed to have common copy number values.

Feature extraction can be considered to be another aspect of “preprocessing”, and as for other preprocessing problems, the statistical community at large has been a little slow to get involved in providing solutions in spite of its clearly quantitative nature. If done well, feature extraction can be an efficient strategy to reduce dimensionality, simplify the data, and focus inference on quantities that are most readily biologically interpretable. However, it is essential that it be done effectively and efficiently, since any biological information in the raw data not contained in the extracted summaries are lost to subsequent analysis. Feature extraction approaches can be much more effective when they are based upon key statistical principles including regularization and unified modeling, which lead to greater efficiency by borrowing strength (i.e. combining information) across subjects, replicates, or genomic regions that are similar to each other. The methods summarized below include feature extraction methods devised by statisticians that either explicitly or implicitly utilize these fundamental principles.

4.1 Statistical Methods for Microarray Preprocessing: Loess Normalization, dChip and RMA

The early 2000’s were characterized by increasing use of microarrays to measure genome-wide gene expression, first on custom cDNA-based spotted arrays and then automated oligonucleotide arrays made by Affymetrix. Empirical investigations of these data revealed various sources of technical variability that sometimes were strong enough to dominate biological variability, including dye bias on spotted arrays, variable probe binding affinities, and general array-specific effects that made combining information across arrays challenging. A number of statisticians rose to these challenges and developed statistical normalization tools that set the field and became the standard tools used by nearly all molecular biologists, as can be seen by the corresponding papers’ high citation counts.

Spotted microarrays typically measured gene expression for pairs of samples, with one sample’s intensities measured by a Cy3 (green) dye and the other measured by a Cy5 (red) dye. Empirical analyses revealed a dye bias, with Cy3 yielding systematically higher values than Cy5 for the same mRNA abundance, yet this relationship did not appear to be linear across the dynamic range. To adjust for this factor, Dudoit et al. (2002) (1730 citations, Google Scholar) developed a robust local linear nonparametric smoother that could be

applied to each array to adjust out this dye effect. This strategy became standard practice, and these types of loess smoothers became a standard normalization tool for all kinds of gene expression data, including non-paired oligonucleotide arrays for which each sample was normalized either to a reference sample (Li and Wong, 2001a) (1115 citations, Google Scholar) or to all others using a pairwise approach (Bolstad et al., 2003) (6124 citations, Google Scholar). This strategy may be the most impactful practical application of the nonparametric smoothing methods developed in the statistical literature starting in the 1980's.

Affymetrix was the first company to mass produce oligonucleotide microarrays, which allowed any investigator to survey whole genome gene expression whether or not their university had its own core facility. To gain more reliable estimates of gene expression than cDNA spotted arrays that typically contained a single probe, Affymetrix included 11–20 probes of 25 base pairs of length for each gene since statistical principles suggested that the averaging over probes should increase efficiency. Recognizing that not all probes have the same binding affinity, for each probe (*perfect match PM*) a second *mismatch probe* (MM) was added for which the 13th base pair was switched. In the software shipped with their arrays, they quantified gene expression by taking a simple average of the differences of $PM - MM$ for all probes for a given array, a method they called *AvDiff*.

The empirical analyses by careful statisticians revealed some problems not handled by this approach, including heteroscedasticity across probes with more abundant probes having greater variability, the presence of outlying probes, samples, or individual observations, and differential binding affinities for different probes that were not adequately handled by the mismatch probes. First, Li and Wong (2001b) (3384 citations, Google Scholar) developed a method *Model-Based Expression Index (MBEI)* distributed as part of the dCHIP software package that used a statistical model to adjust for probe-specific binding affinities, which were estimated by borrowing strength across multiple arrays. For a given sample i and gene, let PM_{ij} and MM_{ij} be the measured perfect match and mismatch expression for probe j , their model was $PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0, \sigma^2)$ whereby θ_i was the gene expression for sample i and ϕ_j the binding affinity effect for probe j with $\sum_j \phi_j^2 = J$ with J is the number of probes for the gene. This model was fitted using alternating least squares applied after performing a loess-based normalization to a reference array, and using an iterative outlier filtering algorithm to remove outlying probes, arrays, and observations, with missing data principles allowing the estimation of the gene expression values θ_i . This method was shown to effectively extend the effective lower detection limit for gene expression.

In their package Robust Multiarray Analysis (*RMA*), Irizarry et al. (2003a) (4161 citations, Google Scholar) improved upon this approach by modeling the log transformed gene expressions, which adjusted for the heteroscedasticity in the data and allowed the fitting of the multiplicative probe affinities in a linear model framework. For robustness they fit their model using robust linear filtering instead of least squares, and adjusted for additive background through transformation and explored array-specific normalization using various approaches (Bolstad et al., 2003), with pairwise nonparametric loess and quantile normalization strategies seeming to work best. They compared their approach with *AvDiff*

and *MBEI*, and found their approach had significant advantages (Irizarry et al., 2003b) in terms of bias, variance, model fit, and detection of known differential expression from spike-in experiments.

These statistical model-based microarray preprocessing packages have become the status quo for preprocessing and are still widely used today.

4.2 Peak Detection for Proteomics

Feature extraction is the predominant strategy for analyzing high throughput proteomics data. For mass spectrometry (MS) data, feature extraction involves detecting peaks on the spectra and then obtaining semi-quantitative measures for each peak by either computing the area under the peak or taking the maximum peak intensity. For 2DGE data, feature extraction involves detecting spots on the gel images, and then quantifying each spot by either taking the volume under the spot within defined spot boundaries or taking the maximum intensity within the spot region. For LC-MS data, feature extraction is sometimes done by summing counts of peptides that map to a given protein. Typically, there is some signal-to-noise (S/N) threshold that must be exceeded for a region to be selected as a feature. For simplicity, we refer to all proteomic objects as “spectra” and all proteomic features as “peaks” for the remainder of this section, although the principles apply across these technologies.

A number of different feature extraction approaches are available in the current literature and in commercial software packages. Until recently, most methods performed detection on individual spectra or gel images, and then matched results across individuals to produce an $n \times p$ matrix of quantifications for p proteins for each of n individuals in the sample. This approach has a number of key weaknesses. It leads to missing data when a given peak does not have a corresponding detected peak for all spectra, and leads to many types of errors, including peak detection errors, peak matching errors, and peak boundary estimation errors that all worsen considerably as the number of spectra increases. These problems are partially responsible for limiting the impact of high-throughput proteomics on biomedical science (Clark and Gutstein, 2008).

The fundamental problem with this strategy is that it uses a piecemeal approach that does not efficiently integrate the information present in the data. Alignment is only done after peak or spot detection, so does not make use of the spatial information in the raw spectra or gels that might lead to improved registration. Peak detection is only done on individual spectra, while ignoring information about whether there appears to be a corresponding peak present in other replicate spectra. If a given potential feature is apparent but near the S/N threshold, knowledge of whether there appears to be a peak at this location in other spectra may help inform the decision of whether it should be flagged as a feature or noise. The detection of peak boundaries from individual spectra also ignores information from other spectra that could be used to refine the boundaries for those features.

Against this backdrop, we developed peak detection methods for MS data (Cromwell, Coombes et al. (2005); Morris et al. (2005), 312 and 324 citations, Google Scholar) and spot detection algorithms for 2DGE data (*Pinnacle*, Morris et al. (2008a)) that utilize

fundamental statistical principles to more efficiently borrow strength within and between spectra, and thus produce substantially improved results. First, spectra are registered to each other in a way that borrows information spatially through their local smoothness properties (Dowsey et al., 2008). Second, rather than detecting peaks on individual spectra, peak detection is performed on the average spectrum, computed by taking the point-wise average across the registered spectra. As demonstrated by Morris et al. (2005) and Morris et al. (2008a), this approach leads to more accurate peak detection, since the averaging reduces the noise by a factor of $\sqrt{n}$ while reinforcing signals present on many individual spectra, which leads to increased signal-to-noise ratios for peaks present in many spectra. Since this approach effectively borrows strength across spectra, the detection accuracy actually increases as the analysis includes more spectra. This is in contrast to standard approaches for which larger numbers of spectra produce increasing propagation of errors (and I have heard many proteomic researchers give this as a reason for running small, and thus underpowered, studies with few samples). Third, wavelet thresholding is used to adaptively denoise the mean spectrum, and its adaptive properties allow the removal of many spurious peaks while retaining the dominant ones. Fourth, peaks are quantified for each individual spectrum by taking the local maximum within a neighborhood of the peak location on the mean spectrum. This ensures there is no missing data, and as shown by Morris et al. (2008a, 2010), this estimation of peak intensity using a local maximum precludes the need to estimate peak boundaries, which leads to more precise and reliable peak quantifications.

These papers have been highly cited and software made freely available for implementing the corresponding approaches. Also, since the publishing of these papers, the commercial software packages for preprocessing proteomics data have incorporated many of these principles and as a result improved their performance.

4.3 Segmentation of DNA copy number data

As alluded to in Section 2.2, array CGH (aCGH) based methods provide a high-resolution view of the DNA-based copy number changes across the whole genome. The resulting data consist of log fluorescence intensity ratios of test to reference samples for specific markers along with the genomic/chromosomal co-ordinates. In an idealized scenario where all of the cells in a disease sample have the same genomic alterations and are uncontaminated by normal cells, the log-ratios would assume specific discrete values e.g. for normal probes, $\log_2(2/2) = 0$; single copy losses, $\log_2(1/2) = -1$; single copy gains, $\log_2(3/2) = 0.58$ etc. In this idealized situation, all copy number alterations could be promptly observed from the data – obviating the need for statistical techniques. However, in real applications in disease areas, the log-ratios differ considerably from these expected values for various technical and biological reasons. DNA copy number data are characterized by high noise levels that add random measurement errors to the observations. Also, the DNA material assessed is not completely homogeneous, as there is typically considerable genomic variability across individual disease cells and there also may be contamination with neighboring normal cells. This heterogeneity implies that we actually measure a composite copy number estimate across a mixture of cell types, which tends to attenuate the ratios toward zero. Finally, and most importantly, these genetic aberrations occur in contiguous spatial regions of the

chromosome that often cover multiple markers and can extend up to whole chromosome arms or chromosomes.

In one of the seminal works towards analyzing such data, Olshen et al. (2004) proposed a statistically principled method called circular binary segmentation (CBS) that provides natural way to segment a chromosome into contiguous regions and bypasses parametric modeling of the data. The fundamental novelty of CBS is that, first, it naturally accounts for the genomic ordering of the markers and adaptively determines what segments share common values and thus adaptively borrow strength from nearby markers. This “local” averaging not only reduces noise but also increases precision in detecting the genomic break-points. CBS is freely available as an R-package (DNACopy) and is widely used in numerous papers as evidenced by close to 1400 citations for the original article (Olshen et al., 2004) – as a starting point for analysis of copy number data.

4.4 Integrated Extraction of Genotype and Copy Number

In contrast to aCGH arrays, genome-wide single nucleotide polymorphism (SNP) genotyping platforms or DNA sequencing provide simultaneous information on both genotypes (e.g. allele-specific frequencies) and copy number variants, CNVs (e.g. logR ratios). This allows for a “generalized” genotyping of both SNPs and CNVs simultaneously on a common sample set, with advantages in terms of cost and unified analysis (Yau and Holmes, 2009).

The segmentation methods described in Section 4.3 are useful to detect copy number changes but fail to account for genotype information or the allele-specific information such as B-Allele frequencies that are typically also provided by such arrays. These two metrics are inherently correlated due to the increased number of genotypes available with increasing copy number, and vice-versa. The idea of generalized genotyping is to flexibly model SNP genotyping data in a way that fully exploits the available information by simultaneously modeling structural changes in both the log ratios and the B-allele frequencies. By working in this original two-dimensional feature space, both the distribution and dependency can be used for better copy number inference and detection. This strategy has subsequently been used by a number of recent algorithms (Colella et al., 2007; Wang et al., 2007). In particular, Colella et al. (2007) propose an objective Bayes Hidden-Markov Model which effectively borrows strength between neighboring SNPs by explicitly incorporating distance information as well the genotype information via B-allele frequencies. This framework provides probabilistic quantification of copy number state classifications and significantly improves the accuracy of segmental identification and mapping relative to existing analytical tools that do not explicitly borrow strength between inherent correlated metrics.

4.5 Supercurve for Protein Quantification in RPPA

As introduced in Section 2.3, Reverse-phase protein arrays (RPPA) have been developed to provide quantitative, high-throughput, time- and cost-efficient analysis of proteins and antibodies. Similar to other high-throughput array-based technologies, RPPA data generation undergoes a series of preprocessing steps before a formal downstream analysis. The three main preprocessing steps are background subtraction, quantification and normalization

(Neeley et al., 2009; Zhang et al., 2009). Background correction involves subtraction of baseline or non-specific signals from the foreground intensities. Once the raw intensities from the RPPA slides have been adjusted for background and other spatial trends, the next preprocessing step is to quantify/estimate the concentration of each protein and sample based on the underlying assumption that the intensity of a given spot on the array is proportional to the amount of protein.

The resulting data consists of a series of dilution series for each RPPA slide, so as to ensure at least one spot from the series in the linear range of expression. As with any standard dilution assays, the expression patterns typically follow a sigmoidal curve i.e. highly diluted spots will often have little protein beyond the background level and conversely, undiluted spots will often have much higher protein levels, with saturation occurring as protein levels get beyond a certain point. The key analytical challenge in protein quantification is to appropriately use the information provided by the entire dilution series to estimate the relative protein abundance for each sample.

A variety of methods have been proposed to estimate the protein concentration. Initial methods performed protein quantification one-sample-at-a-time. Neeley et al. (2009) proposed a joint sample method that aggregates information from all samples on an array to estimate the protein concentration. In essence, it allows borrowing strength across all samples such that all samples on a given array contribute to the overall serial dilution curve. They propose a joint estimation model based on a three-parameter logistic curve and estimating the parameters pooling all the information on an array to estimate global parameters. This joint method has several advantages over naive single sample approaches. First, since each slide is probed with a single antibody targeting that protein, the protein expression of the different samples should share common chemical and hybridization profiles. Second, all of the samples can provide information about the baseline and saturation level, as well as the rate of signal increase at each dilution point. Third, estimating parameters using pooled data can yield more accurate estimates with smaller variances and joint estimation can yield estimates with more dynamic range (Neeley et al., 2009). An R package, SuperCurve, developed to use with this joint estimation method is available at <http://bioinformatics.mdanderson.org/Software/OOMPA> – and often serves as the starting point of any RPPA data analyses.

5 Flexible and Unified Modeling

As discussed in Section 5, feature extraction is the predominant analytical strategy for high-throughput genomics and proteomics data. Feature extraction works well when the extracted features contain all of the scientific information in the data, but otherwise may miss key insights since any information not contained in the extracted features is lost to subsequent analysis. Methods that model the raw data in their entirety have potential to capture insights missed by feature extraction approaches. One prevalent way to model the raw data is to use an *elementwise modeling* approach, whereby individual elements in the raw data are modeled independently of each other. Some examples of this strategy would be to independently model RNA probes for expression data, CpG sites for methylation data, individual spectral locations for MS proteomics data, and SNPs for DNA-based data. This

strategy has the advantage of modeling all of the data and being straightforward to implement, with one able to apply any desired statistical model in parallel to the different elements of the object. However, this strategy ignores the correlation structure among the elements, which has several statistical disadvantages. While unbiased, it leads to inefficient estimators, suboptimal inference, and can exacerbate the inherent multiple testing problem.

In recent years, advances in statistical modeling have led to a growing set of tools available to build flexible statistical models. This includes the development of various basis function modeling approaches including splines, various types of wavelets, empirical basis functions like principal components or independent components, and radial basis functions. Other advances include the development of penalized likelihood approaches to induce sparsity and regularize the fitting of models in high dimensional spaces, and the parallel development in the Bayesian community of prior distributions that induce sparsity such as spike-slab, Bayesian Lasso, Normal-Gamma, Horseshoe, Generalized Double Pareto, and Dirichlet-Laplace priors, plus Bayesian nonparametric priors to flexibly estimate distributions. Hierarchical models have become fundamental tools in Bayesian modeling, and are able to capture various levels of structure and variability in complex, high-dimensional data. These tools can be combined together to build flexible methods that can capture the structure of complex data generated by modern technologies, take this structure into account in a unified fashion, and provide inferential results for many important questions of interest. Fast computational methods including various types of stochastic EM algorithms and approximate Bayesian computational approaches like variational Bayes have been developed and enable the development of flexible methods that are fast enough to fit to high-dimensional data.

Flexible modeling can bridge the gap between the extremes of reductionistic feature extraction approaches that can miss information contained in the data and elementwise modeling approaches that model all of the data but sacrifice efficiency and inferential accuracy by ignoring relationships in the data. In addition to gaining efficiency by accounting for various types of correlation structures inherent to the data, these models also can have inferential advantages, yielding inference accounting for all sources of variability in the data, and potentially adjusting for multiple testing. These benefits are best realized by models deemed to realistically capture structure in the data, at least empirically, so care should be taken to assess model fit when using flexible modeling approaches.

Below we describe a few methods that attempt to flexibly model high-dimensional genomics data while avoiding feature extraction, and were demonstrated to capture biological information missed by some commonly-used feature extraction approaches. While not commonly used in practice for high-throughput genomics data, we believe flexible modeling strategies like these are promising, and should be further pursued and explored by statistical researchers for complex biomedical data like these.

5.1 Flexible Modeling by Functional Regression Methods

One useful class of flexible modeling methods is *functional regression*, which involves regression analyses for which either the response, predictor, or both are functions or images. This can be applied to genome-wide data including methylation and copy number data by

modeling the data as a function of the chromosomal locus, or to proteomics data by modeling mass spectra as spiky functions of m/z values or 2DGE images or LC-MS profiles as image data, which can be viewed as functional data on a two-dimensional domain. To detect differentially expressed regions of the functions, these functions can be modeled as responses and regressed on outcomes of interest, e.g. case vs. control. In these regression models, the regression coefficients are themselves functions defined on the same space as the responses, and so after model fitting differential expression can be assessed by determining for which functional locations the coefficients differ significantly from zero.

As described in Morris (2015), one of the hallmarks of functional regression is to use basis function representations (e.g. splines, wavelets, principal components) and either L1 or L2 penalization to smooth or regularize the resulting functional coefficients. This use of basis function modeling induces a borrowing of strength across nearby measurements within the function which in turn leads to improved efficiency in estimation and inference over elementwise modeling approaches.

Many functional regression methods in existing literature are designed for simple, smooth functions on 1D Euclidean domains, so may not be suitable for high-throughput genomics data. However, we have developed a series of Bayesian methods for functional response regression based on a functional mixed model framework (Morris and Carroll, 2006; Morris et al., 2008b, 2011; Morris, 2012; Zhu et al., 2012; Zhang et al., 2016; Meyer et al., 2016) that are designed for complex, high-dimensional data like these. The functional mixed model is simply a functional response regression model with additional random effect function terms that can be used to account for between-function correlation induced by the experimental design, e.g. for cluster-sampled or longitudinally observed functions, such that it generalizes linear mixed models to the functional response setting.

Our approach for fitting these data is to first represent the observed functions with a lossless or near-lossless basis representation, fit a basis-space version of the functional mixed model using a Markov Chain Monte Carlo (MCMC), and then project these posterior samples back to the original functional space for Bayesian inference to find differentially expressed regions of the function, which can be done while controlling FDR (Morris et al., 2008b) or experimentwise error rate (Meyer et al., 2016). Prior distributions are placed on the basis-space regression coefficients that induce the type of L1 or L2 penalization behavior that leads to appropriately smoothed/regularized functional coefficients. While any lossless or near lossless basis can be used, much of our work has utilized wavelet bases that are well-suited for capturing local features like peaks, spots and change-points that characterize many types of genomics and proteomics data.

This modeling strategy combines together basis function modeling, mixed models or hierarchical models, and sparsity priors to produce a flexible yet scalable model for complex, high dimensional data that has many positive statistical characteristics. The smoothing of the functional regression coefficients induces a borrowing of strength across functional regions that leads to more efficient estimators. The manner of the basis-space modeling induces intra-functional correlation in the residual errors, which is automatically accounted for in the estimation of functional coefficients leading to more efficient estimates

and more accurate inferences. The fully Bayesian model propagates all uncertainties into the final posterior inference, which can adjust for multiple testing using either an FDR or experimentwise error rate based approach.

Finding Differentially Expressed Proteins—Morris et al. (2008b) applied this strategy to MS proteomics data and Morris (2012) compared results with those obtained by *Cromwell*, a feature extraction based approach that detects and quantifies peaks present in the spectra. We found that the functional regression based approach was able to find approximately double the number of differentially expressed proteomic regions, including all of those found by the feature extraction approach and many others, some of which had no corresponding peak detected. Liao et al. (2013) and Liao et al. (2014) demonstrated that this strategy also works and finds protein differences missed by feature extraction approaches for LC-MS data.

Morris et al. (2011) extended this strategy to image data, and applied it to 2DGE data and Morris (2012) compared results with those obtained by *Pinnacle*, a feature extraction based approach that detects and quantifies spots present in the gels. We found that the functional regression based method was able to find nearly all of the results found by the feature extraction based approach, plus approximately 50% more. Many of these novel results were missed by the spot-based approach because of the problem of co-migrating proteins. In 2DGE, the proteins are not perfectly resolved, and so many times a single visual spot contains a convolution of multiple proteins. While spot-based methods will tend to treat this region as a single spot, the functional regression based method is able to detect a differentially expressed protein that is only represented in part of that spot.

For all of these proteomic applications, the basis-space modeling appeared to capture the complex structure of the data well, as data simulated from the functional model look just like real spectra and gels.

Finding Differentially Methylated Regions—Early studies of methylation focused on *CpG islands*, genomic regions containing high frequency of CpG sites, which contain cytosine and guanine connected by a phosphodiester bond. They frequently occur in the promoter region of genes, and have been thought to be the most relevant methylation sites to study. However, recent findings have resulted in a rethinking of this belief. Irizarry et al. (2009) demonstrated that most methylation alterations in colon cancer were not reflected in the CpG island locations, but in regions some distance from the CpG islands, which they coined *CpG Shores*. This was learned by studying DNA methylation on a genome-wide scale, not just restricting to CpG island sites. This suggests that traditional approaches that focus on specific genomic regions such as CpG islands (i.e. a feature extraction approach) are likely to miss important findings and that genome-wide studies of DNA methylation are preferred. This discovery has led to the development of methylation arrays that sample a broader array of CpG sites along the genome, and now to bisulfite sequencing approaches that can survey all CpG sites across the entire genome (Lister et al., 2009).

The collection of genome-wide methylation data raises issues of modeling. In practice, many researchers out of convenience model each CpG site independently (e.g., (Barfield et al.,

2012; Touleimat and Tost, 2012), for example detecting differentially methylated regions (DMRs) as CpG sites with mean methylation levels differing significantly across different conditions. However, given that methylation levels of nearby CpG sites tend to be similar (Leek et al., 2010), this elementwise modeling approach is inefficient as it ignores the correlation structure in the data. Jaffe et al. (2012) use *post hoc* loess smoothing to account for correlation in the data and gain further efficiency, and Lee and Morris (2015) apply Bayesian functional mixed models to detect DMRs. Through simulations and real examples, Lee and Morris (2015) show that the borrowing of strength from the basis function modeling inherent to the functional regression method leads to clearly higher power and lower FDR for discovering DMRs than elementwise methods that model CpG sites independently and ignore their inherent correlation.

5.2 Unified Segmentation Models for Copy Number Data

As mentioned above, one of the hallmarks of genetic variations in cancer is genomic instability of cancerous cells that are manifested as copy number changes across the genome that can be measured using high-resolution and high-throughput assays such as aCGH and SNP arrays (see Section 2.2 for details). The resulting data consist of log fluorescence ratios as a function of the genomic location and provide a cytogenetic representation of the relative DNA copy number variation. Analysis of such data typically involves estimating the underlying copy number state at each genomic location and segmenting regions of the chromosome with similar copy number states. Modeling and inferential challenges of such data include (1) high-dimensionality of the datasets, (2) existence of serial correlations along the genome and (3) multiple assays from a common pool of subjects.

Most methods proceed by modeling a single sample/array at a time, and thus fail to borrow strength across multiple samples to infer shared regions of copy number aberrations (Olshen et al., 2004; Tibshirani and Wang, 2008). Baladandayuthapani et al. (2010) proposed a hierarchical Bayesian approach to address these challenges based on the characterizations of the copy number profiles as functional data i.e. log-ratios as a function of the genomic location – that allows efficient borrowing of strength both within and across arrays to model such data. This approach is based on a multilevel functional mixed effects model that flexibly models not only within subject variability but also allows us to conduct population-level inference to obtain segments of shared copy number changes. The unified Bayesian model uses piece-wise constant functions with random segmentations for the functions that allows determination of optimal segmental rearrangements from the data and more importantly allow incorporation of biological knowledge in the calling of the states of the segments via a hierarchical prior. Applying this method to a well-studied lung cancer dataset, we found many shared regions/genes of interest that are associated with disease progression that were missed by competing simpler approaches. Regions of shared aberrations are of particular importance in cancer genomics – these are key bits of information that can be used to determine subtypes of disease and in to design personalized therapies based on molecular markers, which is one of the most important problems in cancer research today.

6 Structure Learning and Integration

The sheer volume and information-rich nature of the data generated by various high-throughput technologies has pushed the envelope of analytical tools needed to analyze such data. As alluded to in the previous sections, the experimental procedures producing these data and underlying biological principles undergirding them induce a natural higher order organization and structure to these data elements that ideally should be accounted for or included in any downstream modeling endeavors. For example, this includes correlations across genes present in common biological pathways and relationships among measurements from different technological platforms that each contain different biological information based on their molecular resolution level (e.g. DNA, RNA, protein).

These structures are fundamentally ignored by the prevailing piecemeal, multi-step procedures commonly used in practice, presumably out of convenience and lack of available methods. By failing to integrate information effectively, these strategies potentially sacrifice statistical power for making discoveries, and by not propagating inherent correlations throughout inference, they fail to yield optimal inference. By expanding the frontiers of statistical modeling to be able to account for more of this structure, the statistical community has opportunities right now to produce next-generation tools that more efficiently integrate information together, and as a result do a better job of extracting the biological information contained in these rich data.

While by far the most common analytical approaches used are *ad hoc* multi-step algorithmic approaches, model-based approaches if developed to carefully and accurately account for the underlying structure in the data enjoy several inherent advantages. First, they allow full probabilistic formulation of the data generating process. Second, they allow coherent borrowing strength among heterogeneous mixed-scale data sources through appropriate parameterizations. Third, they allow specific inferential questions to be answered through explicit parameterizations. Fourth, they produce uncertainty quantifications and admit natural multiplicity controls. To illustrate these principles, we describe two broad modeling approaches here: *structure learning* and *multi-platform integration*.

6.1 Structure Learning

It is well-established that genes/proteins function in co-ordination within organized modules such as functional or cell signaling pathways or networks (Boehm and Hahn, 2011), for example in cancer to promote or inhibit tumor development. These genes and their corresponding pathways form common modules or networks that regulate various cellular functions. Thus the estimation and incorporation of such modules and networks and their modeling constituents are of great interest for characterizing and understanding the biological mechanisms behind disease development and progression, especially cancer.

Moreover, in the last several years, multiple public and commercial databases have been curated to store the vast amounts of biological knowledge such as signaling, metabolic or regulatory pathways. In general, a *gene class* is defined as a collection of genes defined to be biologically associated given a biological reference based on scientific literature, transcription factor database, expert opinion, or empirical and theoretical evidence. A few of

these include Gene Ontology (GO) (Ashburner et al., 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), MetaCyc (Krieger et al., 2004), Reactome KnowledgeBase (Joshi-Tope et al., 2005), Invitrogen (iPath, www.invitrogen.com) and Cell Signaling Technology (CST) Pathway (www.cellsignal.com). From a statistical viewpoint, models that incorporate these pathway/network structures and combine this outside biological information across genes have been shown to have more statistical power in addition to providing more refined biological interpretations. Following are a few examples.

Gene Set analyses—Broadly, gene set analyses refers set of methods and procedures for integrating the observed experimental data with available agene set information within various scientific contexts (Newton and Wang, 2015). Two broad categories of such methods as reviewed by Newton and Wang include *uniset methods*, that focus on methods for gene sets considered one at a time, ando *multi-set methods* that simultaneously model all of the sets as a unified collection.

A widely used uniset method that relates pathways to a set of experimental data on genes is called gene set enrichment analysis (GSEA, Subramanian et al. (2005)). Briefly, given a list of genes that show some level of significant activity (e.g. differential expression, fold change etc), GSEA computes an “enrichment score” to reflect the degree to which a predefined pathway (using the databases above) is over-represented and then can be used to obtain ranked lists. These procedures are useful starting points to summarize gene expression changes for known biological processes. However, most uniset methods suffer from two drawbacks/challenges as outlined by Newton and Wang. The first is that the set size affects testing power due to imbalance between null and alternate hypothesis (Newton et al., 2007). This imbalance affects how one prioritizes or ranks pertinent gene sets in a given analysis. The second pertains to overlap among different gene sets in their membership, called pleiotropy, which may lead to spurious gene set associations (Bauer et al., 2010; Newton et al., 2012).

Multi-set methods including many model-based methods can alleviate these issues to a certain degree. Examples of multi-set methods include model-based gene set analysis (MGSA) (Bauer et al., 2010, 2011), and multifunctional analysis (MFA) (Wang et al., 2015). Briefly, the major features of model-based statistical approaches are that they bypass the size as well as overlap problems via explicit representations of the gene-level data using latent representations. These multiset methods often involve relatively sophisticated computations and optimizations, and they have demonstrated improved performance over their heuristic/algorithmic uniset counterparts (see Newton and Wang (2015) for further details).

Network and Graphical Models—Graphs and networks provide a natural way of representing the dependency structure among variables. Increasingly, network data are being generated from many scientific areas, especially biology, where large-scale protein-protein interaction and gene regulatory networks are now routinely available. The key scientific hypothesis underpinning the statistical approaches is to look at the system of variables (genes/proteins) as a *whole* rather than individual elements to understand their implicit dependencies. There are typically two key modeling and inferential challenges that underlie

these endeavors. The first is to construct the graph/network based on observed data and second is to use the knowledge to guide models in supervised (e.g regression) and unsupervised (e.g. clustering) settings. These challenges are further accentuated by the fact that typically in many settings the variables far exceed the sample size (a.k.a big n , small p problem).

Graphical models are statistical models that use a graph-based representation to compactly describe probabilistic relationships between the variables, typically genes or proteins. There are two main approaches to estimate the graphs/networks – undirected networks and directed networks, which further incorporate directionality between the edges. In an undirected setting, perhaps the most ubiquitous models are Gaussian graphical models (GGMs) (Cox and Wermuth, 1996) for which the conditional dependencies are encoded by the non-zero entries in the concentration or precision matrix, which is the inverse of covariance matrix of the data. These models provide representations of the conditional independence structure of the multivariate distribution – to develop and infer gene/protein networks. Estimation and application of GGMs have seen a surge in recent years, especially for high-dimensional genomic settings (Meinshausen and Bühlmann, 2006; Friedman et al., 2008; Dobra et al., 2004; Ni et al., 2016). Most of these methods have been widely used in high-dimensional bioinformatics settings, where the primary objective is to induce *sparsity* based on shrinkage and model selection – which can lead to better edge detection with more statistical power and lower false discovery rates. This usually serves as a first step in filtering for empirical structure supported by the data for further downstream experimental validations.

In contrast to the undirected setting, probabilistic network-based approaches, such as Directed Acyclic Graphs (DAG) or Bayesian networks (BN) aim to search through the space of all the possible topological network arrangements while incorporating certain constraints such as directionality (Chen et al., 2006; Myllymäki et al., 2002; Werhli et al., 2006). This has the potential to discover potential casual mechanisms between genes that are not typically accorded by other more naive methods. Various DAG-based methods have been proposed in the literature for use in a variety of contexts. Friedman et al. (2000) developed DAGs from gene expression data using a bootstrap-based approach. Li, Yang, and Xing (2006) constructed DAG-based gene regulatory networks from expression microarray data using linear approaches. Stingo et al. (2010) proposed a DAG-based model to infer microRNA regulatory networks. Recently, Ni et al. (2015) developed an efficient Bayesian method for discovering non-linear edge structures in DAG models which allows the functional form of the relationships between nodes to be non-parametrically determined by the data.

In regression settings, there is a growing set of literature containing methods for conducting variable/feature selection using structured covariates lying on a known graph. These developments mirror the growing recognition that incorporation of supplementary biological information in the analysis of genomic data can be instrumental for improving inference (Pan et al., 2010). Such developments have been aided by a proliferation of genomic databases storing pathway and gene-gene interaction information (Stingo et al., 2011; Li and Zhang, 2010; Shen et al., 2012), and different procedures have been proposed to incorporate available prior information in building structured penalties in a regression model for gene

grouping and selection. Park et al. (2007) attempted to incorporate Gene Ontology pathway information to predict survival time. Some examples of Bayesian regression and variable selection approaches for graph structured covariates include Stingo et al. (2011) and Li and Zhang (2010). Recently, estimation and computational approaches have been developed to generalize graphical model estimation for multi-platform data to infer more integrated networks for get a holistic view of the dependencies (Ha et al., 2015; Ni et al., 2014, 2016). All of these examples demonstrate that the incorporation of external biological information into the modeling leads to better variable selection properties than the more commonly used approaches that ignore such information, and thus borrowing strength from existing resources can lead to more efficient and refined analyses.

6.2 Integromics

A rather nascent but burgeoning field is the field of “integromics” – integrative analysis of multi-platform genomics data. Initial studies in genomics relying on single platform analyses (mostly gene expression- and protein- based) have discovered multiple candidate “druggable” targets especially in cancer such as KRAS mutation in colon and lung cancer (Capon et al., 1982) and BRAF in colorectal, thyroid, and melanoma cancers (Davies et al., 2002). However, it is believed that integrating data across multiple molecular platforms has the potential to discover more co-ordinated changes on a global level (Chin et al., 2011). Integromics espouses the philosophy that a disease is driven by numerous molecular/genetic alterations and the interactions between them, with each type of alteration likely to provide a unique but complementary view of disease progression. This offers a more holistic view of the genomic landscape of a given disease with increased power and lower false discovery rates in detecting important biomarkers (Tyekucheva et al., 2011; Wang et al., 2013), and translating to substantially improved understanding, clinical management and treatment.

The integration of data across diverse platforms has sound biological justifications because of the natural interplay among diverse genomic features. Looking across platforms, attributes at the epigenetic and DNA level such as methylation and copy number variation can affect mRNA expression, which in turn is known to influence clinical outcomes such as progression times and stage of disease through proteins and subsequent post-translational modifications. Figure 1 illustrates some of these inter-platform relationships. Within-platform interactions arise from pathway-based dependencies as well as dependencies based on chromosomal/genomic location. We review some of the recent developments in this area, mostly in the context of cancer, since it is one of the most well-characterized disease-system at different molecular levels. Large scale co-ordinated efforts in cancer include worldwide consortiums such as the International Cancer Genome Consortium (ICGC; icgc.org) and The Cancer Genome Atlas (TCGA; cancergenome.nih.gov), which have collated data over multiple types of cancer on diverse molecular platforms. This has led to proliferation of statistical, bioinformatics and data mining efforts to collectively analyze and model the large volume of data.

Statistically, there are multiple types of data integration methods depending on the scientific question of interest, and the taxonomy can be classified into three broad categories (Kristensen et al., 2014). The first class of methods deal with understanding mechanistic

relationships between different molecular platforms, with the main objective being to delineate cross-platform interactions such as DNA-mRNA, mRNA-protein etc. The second class of methods involves the identification of latent groups of patients or genes using the multi-platform molecular data and can be cast as either a classification (supervised) or clustering (unsupervised) problems. Finally, the third class of methods deals with prediction of an outcome or phenotype (e.g. survival/stage, treatment outcomes) for prospective patients. Some methods focus on one of these categories while others simultaneously consider multiple ones.

Early attempts at integromics involved the sequential analysis of the data from the different platforms in order to understand the biological evolution of disease as opposed to predicting clinical outcome (Fridlyand et al., 2006; Tomioka et al., 2008; Qin, 2008). Briefly, data obtained from one platform are analyzed along with the clinical outcome data, and then a second data platform is subsequently used to clarify or confirm the results obtained from the first platform. For example, Qin (2008) showed that microRNA expression can be used to sort tumors from normal tissues regardless of tumor type. The study then analyzed the relationship between the candidate target genes for the cancer-related microRNAs and mRNA expression and disease status.

More recently, model-based methods have been proposed for which data from multiple platforms are combined into one statistical model. Model-based methods have the advantage of incorporating structural assumptions directly into model-building and several inferential questions can be framed based on appropriate parameterizations. Lanckriet et al. (2004) propose a two stage approach, first computing a kernel representation for data in each platform and subsequently combining kernels across platforms in a classification model. Mo et al. (2013) and Shen et al. (2013) proposed a clustering model “iCluster”, which uses a joint latent variable model to cluster samples into tumor subtypes. Through applications to breast and lung cancer data, iCluster identified potential novel tumor subtypes. Similarly, Lock et al. (2013) proposed an additive decomposition of variation approach consisting of low-rank approximations capturing joint variation across and within platforms, while using orthogonality constraints to ensure that patterns within and across platforms are unrelated. Tyekucheva et al. (2011) proposed a logistic regression model regressing a clinical outcome on covariates across multiple platforms.

iBAG: Integrative Bayesian analysis of genomics data—Recently Wang et al. (2013) introduced integrative Bayesian analysis of genomics data (iBAG), a unified framework for integrating information across genomic, transcriptomic and epigenomic data as well clinical outcomes. iBAG uses a two-component hierarchical model construction: a *mechanistic model* to infer direct effects of different platforms on gene expression, and a *clinical model* that uses this information to associate with a relevant clinical outcome (e.g. survival times). The mechanistic model takes into account the biological relationships between platforms by modeling mRNA gene expression as a linear or nonlinear function of its upstream regulators and decomposes a given gene’s expression into separate components that are regulated by each upstream platform and a residual component that accounts for expression effectors not included in the model. This serves a bi-fold objective: first, it captures the mechanistic dependencies among different platforms modulating the expression

and second, serves as a denoising step for the expression before correlating them with the clinical outcomes. Subsequently, the clinical component hierarchically learns from the mechanistic model by incorporating the platform-specific genomic components and the clinical factors (age, stage, demographics) into one model including multiple genes to find “optimal” integrative signatures associated with the clinical outcomes.

The authors demonstrated that this modeling framework, by statistically borrowing strength across different data sources, allows: (a) better delineation of biological mechanisms between different platforms; (b) increased power to detect important biomarkers of disease progression and (c) increased prediction accuracy for clinical outcomes. This framework has been further generalized to multiple platforms (Jennings et al., 2013), incorporating non-linear dependencies (Jennings et al., 2016), and is in the process of being extended to incorporate pathway-based dependencies and integrate radiology-based imaging data.

These methods exemplify how integrative statistical models can be used to borrow strength from disparate data sources in order to perform more refined analyses that have potential to provide additional insights into the biological mechanisms governing the disease processes that might be missed by the prevalent piecemeal approaches.

7 Conclusions and Future Directions

Statisticians have played a prominent role in bioinformatics, helping develop rigorous design and analysis tools for researchers to use to extract meaningful biological information from the rich treasure trove of multi-platform genomics data. Their deep understanding of the scientific process as well as variability and uncertainty has uniquely equipped them to serve a fundamental role in this venture.

In this paper, we have attempted to summarize this contribution, focusing on four key areas of experimental design and reproducibility, preprocessing, unified modeling, and structure learning and integration. There has been considerable high-impact work done in these areas, and the success and benefit of these statistician-derived methods is driven by the key statistical concepts motivating and underlying them.

One of the key statistical concepts is that unified models that borrow strength across related elements enjoy statistical benefits over piecemeal approaches, leading to more efficient estimation, improved prediction, and greater sensitivity and lower false discovery rates for making discoveries. This borrowing of strength can occur across samples, across measurements within an object (e.g. across probes, spectral locations, genomic locations, or genes in common pathway), across data types, and between data and biological knowledge in the literature. In this paper, we see this concept at work in peak detection on the mean spectrum, incorporation of copy number and B-allele frequency to determine copy number estimates, borrowing of strength across samples to estimate underlying protein abundances, borrowing strength across samples to identify shared genomic copy number aberrations, incorporating pathway information into models, or integrating across platforms using DAGs or hierarchical models that model their natural interrelationships.

This principle is also at work in flexible modeling approaches that borrow strength across nearby observations in functional or image data using basis function modeling and regularization priors, a strategy that has been applied to MS, 2DGE, copy number, and methylation data. The concept of regularization is used when smoothing functional data in normalization of microarrays, when penalizing regression coefficients in high-dimensional regression models, when denoising spectra before performing peak detection, and when segmenting DNA copy number data.

By applying these principles, we can continue to develop efficient methods that can strongly impact the field of bioinformatics moving forward. New technologies are continually being developed and introduced at a rapid rate, and there are many new challenges these data will bring. Our hope is that statisticians will be involved on the front lines of methods development for these technologies as they are introduced, and that we are involved in all aspects of the science including design, preprocessing, and end-stage analysis, not just end-stage analysis.

We acknowledge that some of the genomic platforms featured in this paper comprise older technologies that have been mostly supplanted by newer ones. However, our experience is that while new technologies always bring some new challenges, many of the quantitative issues remain the same. Thus, methods and approaches developed on older platforms have some translational importance to the new ones, at least in terms of key issues and the underlying principles behind effective solutions to them. Our hope is that by elucidating key statistical principles driving some of these methods, we will help stimulate future researchers in finding effective solutions to future challenges.

There are a number of areas where more work is clearly needed and future developments are possible. One key area is in integrative analysis. This field is really just getting started and the scientific community is in dire need of new methods for integrating information across multiple platforms to gain more holistic insights into the underlying molecular biology. These methods must balance statistical rigor in building connections, computational efficiency to scale up to big data settings, and interpretability of results so our collaborators can make sense of them. Also, given the extensive efforts in the biological research community to build up knowledge resources that are freely available online, such the recent large-scale federal efforts for unified databases especially in cancer e.g. NCI Genomic Data Commons (GDC, Grossman et al. (2016)). Hence, the statistical community needs to find better ways to incorporate this information into the modeling, which can lead to improved predictions and discoveries as well as enhanced interpretability of the results. Given the interdependencies underlying genetic processes, pathway information is one of the most important types of information that we need to better incorporate.

Biology and medicine have moved to a place where big data are becoming ubiquitous in research and even clinical practice. This provides great opportunities for the statistical community to play a fundamental role in pushing the science forward, as we equip other scientists with the tools they need to extract the valuable information they contain.

Acknowledgments

This work has been supported by grants from the National Cancer Institute (R01-CA178744, P30-CA016672, R01-CA160736, R01-CA194391) and the National Science Foundation (1550088, 1463233).

References

- Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*. 2000; 403(6769):503–511. [PubMed: 10676951]
- Alwine JC, Kemp DJ, Stark GR. Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes. *Proceedings of the National Academy of Sciences*. 1977; 74(12):5350–5354.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*. 2000; 25(1):25–29. [PubMed: 10802651]
- Augustin CK, Yoo JS, Potti A, Yoshimoto Y, Zipfel PA, Friedman HS, Nevens JR, Ali-Osman F, Tyler DS. Genomic and molecular profiling predicts response to temozolomide in melanoma. *Clinical Cancer Research*. 2009; 15:502–510. [PubMed: 19147755]
- Baggerly KA, Coombes KR. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*. 2010; 3(4): 1309–1334.
- Baggerly KA, Edmonson SR, Morris JS, Coombes KR. High-resolution serum proteomic patterns for ovarian cancer detection. *Endocrine-Related Cancer*. 2004a; 11(4):583–584. [PubMed: 15613439]
- Baggerly KA, Morris JS, Coombes KR. Reproducibility of selditof protein patterns in serum: Comparing data sets from different experiments. *Bioinformatics*. 2004b; 20:777–785. [PubMed: 14751995]
- Baggerly KA, Coombes KR, Morris JS. Bias, randomization, and ovarian proteomic data: A reply to ? producers and consumers. *Cancer Informatics*. 2005a; 1(1):9–14. [PubMed: 19305627]
- Baggerly KA, Morris JS, Edmonson SR, Coombes KR. Signal in noise: Evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *Journal of the National Cancer Institute*. 2005b; 97(4):307–309. [PubMed: 15713966]
- Baladandayuthapani V, Ji Y, Talluri R, Nieto-Barajas LE, Morris JS. Bayesian random segmentation models to identify shared copy number aberrations for array cgh data. *Journal of the American Statistical Association*. 2010; 105(492):1358–1375. [PubMed: 21512611]
- Baladandayuthapani V, Talluri R, Ji Y, Coombes KR, Lu Y, Hennessy BT, Davies MA, Mallick BK. Bayesian sparse graphical models for classification with application to protein expression data. *The Annals of Applied Statistics*. 2014; 8(3):1443. [PubMed: 26246866]
- Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Research*. 2011; 21(3):381–395. [PubMed: 21321607]
- Barfield RT, Kilaru V, Smith AK, Conneely KN. CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics*. 2012; 28(9):1280–1281. [PubMed: 22451269]
- Bauer S, Gagneur J, Robinson PN. Going bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Research*. 2010; gkq045.
- Bauer S, Robinson PN, Gagneur J. Model-based gene set analysis for bioconductor. *Bioinformatics*. 2011; 27(13):1882–1883. [PubMed: 21561920]
- Begley CG, Ellis L. Drug development: Raise standards for preclinical cancer research. *Nature*. 2012; 483:531–533. [PubMed: 22460880]
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *JRSS-B*. 1995; 57:289–300.
- Beroukhim R, Lin M, Park Y, Hao K, Zhao X, Garraway LA, Fox EA, Hochberg EP, Mellinghoff IK, Hofer MD, et al. Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide snp arrays. *PLoS Comput Biol*. 2006; 2(5):e41. [PubMed: 16699594]

- Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*. 2001; 98(24):13790–13795.
- Bibikova JL, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, Gunderson KL. Genome-wide dna methylation profiling using infinium assay. *Epigenetics*. 2009; 1:177–200.
- Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB, Shen R. High density dna methylation array with single cpg site resolution. *Genomics*. 2011; 98(4):288–295. [PubMed: 21839163]
- Bild AH, Chang JT, Johnson WE, Piccolo SR. A field guide to genomics research. *PLoS Biol*. 2014; 12(1):e1001744. [PubMed: 24409093]
- Bird AP, Taggart MH, Nicholls RD, Higgs DR. Non-methylated cpg-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *The EMBO journal*. 1987; 6(4):999–1004. [PubMed: 3595568]
- Blanchard A, Kaiser R, Hood L. High-density oligonucleotide arrays. *Biosensors and bioelectronics*. 1996; 11(6):687–690.
- Boehm JS, Hahn WC. Towards systematic functional characterization of cancer genomes. *Nature Reviews Genetics*. 2011; 12(7):487–498.
- Bolstad BM, Irizarry RA, Ashard M, Speed TP. A comparison of normlization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003; 19(2):185–193. [PubMed: 12538238]
- Bonato V, Baladandayuthapani V, Broom BM, Sulman EP, Aldape KD, Do KA. Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*. 2011; 27(3):359–367. [PubMed: 21148161]
- Bonnefoi H, Potti A, Delorenzi M, Mauriac L, Campone M, Tubianahulin M, Petit T, Rouanet P, Jassem J, Blot E, Becette V, Farmer P, Andre S, Acharya CR, Mukherjee S, Cameron D, Bergh J, Nevins JR, Iggo RD. Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: A substudy of the eortc 10994/big 00-01 clinical trial. *Lancet Oncology*. 2007; 8:1071–1078. [PubMed: 18024211]
- Brown PO, Botstein D. Exploring the new world of the genome with dna microarrays. *Nature Genetics*. 1999; 21:33–37. [PubMed: 9915498]
- Bueno-de Mesquita JM, van Harten WH, Retel VP, van't Veer LJ, van Dam FS, Karsenberg K, Douma KF, van Tinteren H, Peterse JL, Wesseling J, et al. Use of 70-gene signature to predict prognosis of patients with nodenegative breast cancer: a prospective community-based feasibility study (raster). *The Lancet Oncology*. 2007; 8(12):1079–1087. [PubMed: 18042430]
- Capon DJ, Seeburg PH, McGrath JP, Hayflick JS, Edman U, Levinson AD, Goeddel DV. Activation of ki-ras2 gene in human colon and lung carcinomas by two different point mutations. *Nature*. 1982; 304(5926):507–513.
- Cardoso F, Van't Veer L, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ. Clinical application of the 70-gene profile: the mindact trial. *Journal of Clinical Oncology*. 2008; 26(5):729–735. [PubMed: 18258980]
- Chen X, Chen M, Ning K. Bnarray: an r package for constructing gene regulatory networks from microarray data by using bayesian network. *Bioinformatics*. 2006; 22(23):2952–2954. [PubMed: 17005537]
- Chin L, Hahn WC, Getz G, Meyerson M. Making sense of cancer genomic data. *Genes & Development*. 2011; 25(6):534–555. [PubMed: 21406553]
- Clark BN, Gutstein HB. The myth of automated, high-throughput two-dimensional gel analysis. *Proteomics*. 2008; 8:1197–1203. [PubMed: 18283661]
- Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*. 2007; 35(6):2013–2025. [PubMed: 17341461]
- Collas P. The current state of chromatin immunoprecipitation. *Molecular Biotechnology*. 2010; 45(1): 87–100. [PubMed: 20077036]

- Collins FS, Tabak LA. Policy: Nih plans to enhance reproducibility. *Nature*. 2014; 505(7485):612–613. [PubMed: 24482835]
- Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, Hung MC, Kuerer HM. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*. 2005; 5:4107–4117. [PubMed: 16254928]
- Cox, DR., Wermuth, N. Multivariate dependencies: Models, analysis and interpretation. Vol. 67. CRC Press; 1996.
- Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, et al. Mutations of the braf gene in human cancer. *Nature*. 2002; 417(6892):949–954. [PubMed: 12068308]
- DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*. 1997; 278(5338):680–686. [PubMed: 9381177]
- Discover. Discover. 2007 Jan. The top 6 genetics stories of 2006.
- Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*. 2004; 90(1):196–212. URL <http://www.sciencedirect.com/science/article/B6WK9-4C604WK-1/2/9a861453b1df438db4cff4e718f94246>. Special Issue on Multivariate Methods in Genomic Data Analysis. DOI: 10.1016/j.jmva.2004.02.009
- Dowsey A, Dunn M, Yang G. Automated image alignment for 2d gel electrophoresis in a high-throughput proteomics pipeline. *Bioinformatics*. 2008; 24:950–957. [PubMed: 18310057]
- Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*. 2002; 12:111–139.
- Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*. 2004; 99:96–104.
- Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*. 1998; 95(25):14863–14868.
- Fridlyand J, Snijders AM, Ylstra B, Li H, Olshen A, Segraves R, Dairkee S, Tokuyasu T, Ljung BM, Jain AN, et al. Breast tumor copy number aberration phenotypes and genomic instability. *BMC cancer*. 2006; 6(1):1. [PubMed: 16390557]
- Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008; 9(3):432–441. <http://biostatistics.oxfordjournals.org/cgi/content/abstract/9/3/432>. DOI: 10.1093/biostatistics/kxm045 [PubMed: 18079126]
- Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*. 2000; 7(3–4):601–620. [PubMed: 11108481]
- Fuentes M. Reproducible research in jasa. *AmStat News*. 2016 Jul 1.
- Gillespie D, Spiegelman S. A quantitative assay for dna-rna hybrids with dna immobilized on a membrane. *Journal of Molecular Biology*. 1965; 12(3):829–842. [PubMed: 4955314]
- Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*. 2016; 375(12):1109–1112. URL <http://dx.doi.org/10.1056/NEJMp1607591>. DOI: 10.1056/NEJMp1607591 [PubMed: 27653561]
- Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot B, Morris J, Simon I, Gerster S, Fessler E, de Sousa A, Melo F, Missiaglia E, Ramay H, Barras D, Homicsko K, Maru D, Manyam G, Broom B, Boige V, Laderas T, Salazar R, Gray J, Tabernero J, Bernards R, Friend S, Laurent-Puig P, Medema J, Sadanandam A, Wessels L, Delorenzi M, Kopetz S, Vermeulen L, Tejpar S. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*. 2015; 21(11):1350–1356.
- Ha MJ, Baladandayuthapani V, Do KA. Dingo: differential network analysis in genomics. *Bioinformatics*. 2015; 31(21):3413–3420. [PubMed: 26148744]
- Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CM, Beyene J. Data integration in genetics and genomics: methods and challenges. *Human Genomics and Proteomics*. 2009; 1(1)

- Hennessy B, Lu Y, Gonzalez-Angulo A, Carey M, Myhre S, Ju Z, Davies M, Liu W, Coombes K, Meric-Bernstam F, et al. A technical assessment of the utility of reverse phase protein arrays for the study of the functional proteome in non-microdissected human breast cancers. *Clinical Proteomics*. 2010; 6(4):129–151. [PubMed: 21691416]
- Hofner B, Schmid M, Edler L. Reproducible research in statistics: A review and guidelines for the biometrical journal. *Biometrical Journal*. 2016; 58(2):416–427. [PubMed: 26711717]
- Hsu DS, Balakumaran BS, Acharya CR, Vlahovic V, Walters KS, Garman K, Anders C, Riedel RF, Lancaster J, Harpole D, Dressman HK, Nevins JR, Febbo PG, Potti A. Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer. *Journal of Clinical Oncology*. 2007; 25:4350–4357. [PubMed: 17906199]
- Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, Van Noort V. Repeatability of published microarray gene expression analyses. *Nature Genetics*. 2009; 41:149–155. [PubMed: 19174838]
- Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*. 2003a; 31(4):e15. [PubMed: 12582260]
- Irizarry RA, Hobbs B, Collin F, Beazer-Barday YD, Antonelli KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003b; 4(2):249–264. [PubMed: 12925520]
- Irizarry RA, Ladd-Acosta CL, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, Ji H, Potash JB, Sabuncuyan S, Feinberg AP. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*. 2009; 41:178–186. [PubMed: 19151715]
- Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*. 2012; 41(1):200–209. [PubMed: 22422453]
- Jennings EM, Morris JS, Manyam GC, Carroll RJ, Baladandayuthapani V. Bayesian models for flexible integrative analysis of multi-platform genomic data. 2016
- Jennings EM, Morris JS, Carroll RJ, Manyam G, Baladandayuthapani V. Bayesian methods for expression-based integration of various types of genomics data. *EURASIP J. Bioinformatics and Systems Biology*. 2013; 2013:13.
- Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath G, Wu G, Matthews L, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*. 2005; 33(suppl 1):D428–D432. [PubMed: 15608231]
- Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*. 1992; 258(5083):818–821. [PubMed: 1359641]
- Kanehisa M, Goto S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. 2000; 28(1):27–30. [PubMed: 10592173]
- Karp NA, Lilley KS. Maximizing sensitivity for detecting changes in protein expression: Experimental design using minimal cydyes. *Proteomics*. 2005; 5:3105–3115. [PubMed: 16035117]
- Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD. Metacyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*. 2004; 32(suppl 1):D438–D442. [PubMed: 14681452]
- Kristensen VN, Lingjærde OC, Russnes HG, Vøllan HKM, Frigessi A, Børresen-Dale AL. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*. 2014; 14(5):299–313. [PubMed: 24759209]
- Laanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. *Bioinformatics*. 2004; 20(16):2626–2635. [PubMed: 15130933]
- Lee W, Morris J. Identification of differentially methylated loci using wavelet-based functional mixed models. *Bioinformatics*. 2015
- Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*. 2010; 11(10):733–739.

- Li C, Wong W. Model-based analysis of oligonucleotide arrays: model validation, design issues, and standard error approximation. *Genome Biology*. 2001a; 2(8):RESEARCH 0032.
- Li C, Wong W. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Science (USA)*. 2001b; 98:31–36.
- Li F, Zhang NR. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*. 2010; 105(491):1202–1214.
- Li, F., Yang, Y., Xing, E. Inferring regulatory networks using a hierarchical Bayesian graphical Gaussian model. Carnegie Mellon University, School of Computer Science, Machine Learning Department; 2006.
- Liao H, Moschidis E, Riba-Garcia I, Zhang I, Unwin R, Morris J, Graham J, Dowsey A. A new paradigm for clinical biomarker discovery and screening with mass spectrometry based on biomedical image analysis principles. *IEEE International Symposium on Biomedical Imaging*. 2014
- Liao, L., Moschidis, E., Riba-Garcia, I., Unwin, R., Dunn, W., Morris, J., Graham, J., Dowsey, A. A workflow for novel image-based differential analysis of lc-ms experiments. *Proceedings of 61st ASMS Conference on Mass Spectrometry and Allied Topics*; 2013.
- Liotta LA, Lowenthal M, Mehta A, Conrades TP, Veenstra TD, Fishman DA, Petricoin EF III. Importance of communication between producers and consumers of publicly available experimental data. *Journal of the National Cancer Institute*. 2005; 97(4):310–314. [PubMed: 15713967]
- Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. Human dna methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009; 462:315–322. [PubMed: 19829295]
- Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The Annals of Applied Statistics*. 2013; 7(1):523. [PubMed: 23745156]
- Lucito R, West J, Reiner A, Alexander J, Esposito D, Mishra B, Powers S, Norton L, Wigler M. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Research*. 2000; 10(11):1726–1736. [PubMed: 11076858]
- Mallick, BK., Gold, DL., Baladandayuthapani, V. *Front Matter*. Wiley Online Library; 2009.
- Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, Patil N, Wolff RK, Chee MS, Reid BJ, Lockhart DJ. Genome-wide detection of allelic imbalance using human snps and high-density dna arrays. *Genome Research*. 2000; 10(8):1126–1137. [PubMed: 10958631]
- Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*. 2006; 34(3):1436–1462. URL <http://www.jstor.org/stable/25463463>.
- Meyer M, Coull B, Versace F, Cinciripini P, Morris J. Bayesian function-on-function regression for multi-level functional data. *Biometrics*. 2016; 71(3):563–574.
- Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*. 2013; 110(11):4245–4250.
- Morris JS, Coombes K, Kooman J, Baggerly K, Kobayashi R. Feature extraction and quantification for mass spectrometry data in biomedical applications using the mean spectrum. *Bioinformatics*. 2005; 21:1764–1775. [PubMed: 15673564]
- Morris JS, Clark B, Gutstein H. Pinnacle: A fast, automatic and accurate method for detecting and quantifying protein spots in 2-dimensional gel electrophoresis data. *Bioinformatics*. 2008a; 24:529–536. [PubMed: 18194961]
- Morris JS, Clark BN, Wei W, Gutstein HB. Evaluating the performance of new approaches to spot quantification and differential expression in 2-dimensional gel electrophoresis studies. *Journal of Proteome Research*. 2010; 9(1):595–604. [PubMed: 19919108]
- Morris JS, Brown PJ, Herrick RC, Baggerly KA, Coombes KR. Bayesian analysis of mass spectrometry data using wavelet-based functional mixed models. *Biometrics*. 2008b; 12:479–489.

- Morris J. Statistical methods for proteomic biomarker discovery using feature extraction or functional data analysis approaches. *Statistics and its Interface*. 2012; 5(1):117–136. [PubMed: 23814640]
- Morris J. Functional regression. *Annual Review of Statistics and its Application*. 2015; 2:321–359.
- Morris J, Carroll R. Wavelet-based functional mixed models. *J R Statist Soc B*. 2006; 68(2):179–199.
- Morris J, Baladandayuthapani V, Herrick R, Sanna P, Gutstein H. Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *The Annals of Applied Statistics*. 2011; 5:894–923. [PubMed: 22408711]
- Muller P, Parmigiani G, Robert C, Rousseau J. Optimal sample size for multiple testing: The case of gene expression microarrays. *Journal of the American Statistical Association*. 2004; 99(468): 990–1001.
- Myllymäki P, Silander T, Tirri H, Uronen P. B-course: A web-based tool for bayesian and causal data analysis. *International Journal on Artificial Intelligence Tools*. 2002; 11(03):369–387.
- Neeley ES, Kornblau SM, Coombes KR, Baggerly KA. Variable slope normalization of reverse phase protein arrays. *Bioinformatics*. 2009; 25(11):1384–1389. URL <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/11/1384>. DOI: 10.1093/bioinformatics/btp174 [PubMed: 19336447]
- Newton MA, Wang Z. Multiset statistics for gene set analysis. *Annual Review of Statistics and its Application*. 2015; 2:95–111.
- Newton MA, Noueiry A, Sarker D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics*. 2004; 5(2):155–176. [PubMed: 15054023]
- Newton MA, Quintana FA, Den Boon JA, Sengupta S, Ahlquist P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics*. 2007:85–106.
- Newton MA, He Q, Kendzierski C. A model-based analysis to infer the functional content of a gene list. *Statistical Applications in Genetics and Molecular Biology*. 2012; 11(2)
- Ni Y, Stingo FC, Baladandayuthapani V. Integrative Bayesian network analysis of genomic data. *Cancer Informatics*. 2014; 13(Suppl 2):39. [PubMed: 25288878]
- Ni Y, Stingo FC, Baladandayuthapani V. Bayesian nonlinear model selection for gene regulatory networks. *Biometrics*. 2015; 71(3):585–595. [PubMed: 25854759]
- Ni Y, Stingo FC, Baladandayuthapani V. Sparse multi-dimensional graphical models: A unified Bayesian framework. *Journal of the American Statistical Association*. 2016 (to appear).
- O’Farrell PH. High-resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry*. 1975; 250:4007–4021. [PubMed: 236308]
- Olshen AB, Venkatraman E, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*. 2004; 5(4):557–572. [PubMed: 15475419]
- Pan W, Xie B, Shen X. Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*. 2010; 66(2):474–484. [PubMed: 19645699]
- Park MY, Hastie T, Tibshirani R. Averaged gene expressions for regression. *Biostatistics*. 2007; 8(2): 212–227. [PubMed: 16698769]
- Paweletz C, Charboneau L, Bichsel V, Simone N, Chen T, Gillespie J, Emmert-Buck M, Roth M, Petricoin E, Liotta L. Reverse phase protein microarrays which capture disease progression show activation of prosurvival pathways at the cancer invasion front. *Oncogene*. 2001; 20(16):1981–1989. [PubMed: 11360182]
- Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor S. Light-generated oligonucleotide arrays for rapid dna sequence analysis. *Proceedings of the National Academy of Sciences*. 1994; 91(11):5022–5026.
- Peng R. Reproducible research and *Biostatistics*. *Biostatistics*. 2009; 10(3):405–408. [PubMed: 19535325]
- Petricoin EFI, Fishman DA, Conrads TP, Veenstra TD, Liotta LA. Proteomic pattern diagnostics: Producers and consumers in the era of correlative science. comment on sorace and zhan. *BMC Bioinformatics*. 2004

- Petricoin EFI, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mils GB, Simone C, Fishman DA, Kohn EC, Liotta LA. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*. 2002; 359:572–577.
- Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*. 2005; 37:S11–S17. [PubMed: 15920524]
- Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, et al. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*. 1998; 20(2):207–211. [PubMed: 9771718]
- Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO. Genome-wide analysis of dna copy-number changes using cdna microarrays. *Nature Genetics*. 1999; 23(1):41–46. [PubMed: 10471496]
- Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D, Marks J, Berchuck A, Ginsburg GS, Febbo P, Lancaster J, Nevins JR. Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine*. 2006; 12:1294–1300.
- Qin L-X. An integrative analysis of microrna and mrna expression-a case study. *Cancer Informatics*. 2008;6.
- Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*. 2001; 98(26):15149–15154.
- Ramsay, J., Silverman, B. *Functional Data Analysis*. Springer-Verlag; New York: 1997.
- Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*. 1995; 270(5235):467–470. [PubMed: 7569999]
- Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW. Microarrays: biotechnology's discovery platform for functional genomics. *Trends in Biotechnology*. 1998; 16(7):301–306. [PubMed: 9675914]
- Schuster SC. Next-generation sequencing transforms today's biology. *Nature Methods*. 2008; 5(1):16–18. [PubMed: 18165802]
- Seidel C. Introduction to DNA microarrays. *Analysis of Microarray Data: A Network-based Approach*. 2008; 1:1.
- Shalon D, Smith SJ, Brown PO. A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. *Genome Research*. 1996; 6(7):639–645. [PubMed: 8796352]
- Shen R, Wang S, Mo Q. Sparse integrative clustering of multiple omics data sets. *The Annals of Applied Statistics*. 2013; 7(1):269. [PubMed: 24587839]
- Shen X, Huang HC, Pan W. Simultaneous supervised clustering and feature selection over a graph. *Biometrika*. 2012; 99(4):899–914. [PubMed: 23843673]
- Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, et al. Assembly of microarrays for genome-wide measurement of dna copy number. *Nature Genetics*. 2001; 29(3):263–264. [PubMed: 11687795]
- Sorace J, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics*. 2003;4. [PubMed: 12549981]
- Stingo FC, Chen YA, Vannucci M, Barrier M, Mirkes PE. A Bayesian graphical modeling approach to microRNA regulatory network inference. *The Annals of Applied Statistics*. 2010; 4(4):2024–2048. [PubMed: 23946863]
- Stingo FC, Chen YA, Tadesse MG, Vannucci M. Incorporating biological information into linear models: A bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics*. 2011; 5(3)
- Stodden V, Guo P, Ma Z. Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLOS One*. 2013; 8(6):e67111. [PubMed: 23805293]
- Storey JD. A direct approach to false discovery rate. *JRSS-B*. 2002; 64:479–498.
- Storey JD. A positive false discovery rate: a bayesian interpretation and the q-value. *Annals of Statistics*. 2003; 31:2013–2035.

- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*. 2005; 102(43):15545–15550.
- Tibes R, Qiu Y, Lu Y, Hennessy B, Andreeff M, Mills GB, Kornblau SM. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular Cancer Therapeutics*. 2006; 5(10):2512–2521. [PubMed: 17041095]
- Tibshirani R, Wang P. Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*. 2008; 9(1):18–29. [PubMed: 17513312]
- Tomioka N, Oba S, Ohira M, Misra A, Fridlyand J, Ishii S, Nakamura Y, Isogai E, Hirata T, Yoshida Y, et al. Novel risk stratification of patients with neuroblastoma by genomic signature, which is independent of molecular signature. *Oncogene*. 2008; 27(4):441–449. [PubMed: 17637744]
- Touleimat N, Tost J. Complete pipeline for infinium human methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*. 2012; 4(3):325–341. [PubMed: 22690668]
- Tukey, JW. *Exploratory Data Analysis*. Addison-Wesley Publishing Company; 1977.
- Tyekucheva S, Marchionni L, Karchin R, Parmigiani G. Integrating diverse genomic data using gene sets. *Genome Biology*. 2011; 12(10):1–14.
- Visser LE, de Vries BB, Veltman JA. Genomic microarrays in mental retardation: from copy number variation to gene, from research to diagnosis. *Journal of Medical Genetics*. 2010; 47(5):289–297. [PubMed: 19951919]
- Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*. 2007; 17(11):1665–1674. [PubMed: 17921354]
- Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, Do KA. ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*. 2013; 29(2):149–159. [PubMed: 23142963]
- Wang Z, He Q, Larget B, Newton MA, et al. A multi-functional analyzer uses parameter constraints to improve the efficiency of model-based geneset analysis. *The Annals of Applied Statistics*. 2015; 9(1):225–246.
- Werhli AV, Grzegorzczak M, Husmeier D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*. 2006; 22(20):2523–2531. [PubMed: 16844710]
- Yau C, Holmes C. Cnv discovery using snp genotyping arrays. *Cytogenetic and Genome Research*. 2009; 123(1–4):307–312.
- Zhang L, Baladandayuthapani V, Baggerly K, Czerniak B, Morris J. Functional car models for spatially correlation high-dimensional functional data. *Journal of the American Statistical Association*. 2016 to appear.
- Zhang L, Wei Q, Mao L, Liu W, Mills GB, Coombes K. Serial dilution curve: a new method for analysis of reverse phase protein array data. *Bioinformatics*. 2009; 25(5):650–654. URL <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/5/650>. DOI: 10.1093/bioinformatics/btn663 [PubMed: 19176552]
- Zhu H, Brown P, Morris J. Robust classification of functional and quantitative image data using functional mixed models. *Biometrics*. 2012; 68:1260–1268. [PubMed: 22670567]

Figure 1.
Illustration of Types of Multi-platform Genomics Data and Their Interrelationships

Figure 2.

Heatmap of Ovarian Cancer Data: Heatmap of mass spectra from 216 samples in Petricoin et al. (2002) run on Ciphergen H4 ProteinChip (top) and Ciphergen WCX2 (bottom).