The Promise of Genomic Studies on Human Diseases: From Basic Science to Clinical Application

Guest Editors: Lam C. Tsoi, Bethany Wolf, and Y. Ann Chen
The Promise of Genomic Studies on Human Diseases: From Basic Science to Clinical Application
The Promise of Genomic Studies on Human Diseases: From Basic Science to Clinical Application

Guest Editors: Lam C. Tsoi, Bethany Wolf, and Y. Ann Chen
Editorial Board

Jacques Camonis, France
Prabhakara V. Choudary, USA
Martine A. Collart, Switzerland
Marco Gerdol, Italy
Soraya E. Gutierrez, Chile
M. Hadzopoulou-Cladaras, Greece
Sylvia Hagemann, Austria
Henry Heng, USA
Eivind Hovig, Norway
Giuliana Napolitano, Italy
Ferenc Olasz, Hungary
Elena Pasyukova, Russia
Graziano Pesole, Italy
Giulia Piaggio, Italy
Mohamed Salem, USA
Brian Wigdahl, USA
Jinfa Zhang, USA
## Contents

**The Promise of Genomic Studies on Human Diseases: From Basic Science to Clinical Application**  
Lam C. Tsoi, Bethany Wolf, and Y. Ann Chen  
Volume 2017, Article ID 5093167, 2 pages

**A Review of Recent Advancement in Integrating Omics Data with Literature Mining towards Biomedical Discoveries**  
Kalpana Raja, Matthew Patrick, Yilin Gao, Desmond Madu, Yuyang Yang, and Lam C. Tsoi  
Volume 2017, Article ID 6213474, 10 pages

**Integrating Biological Covariates into Gene Expression-Based Predictors of Radiation Sensitivity**  
Vidya P. Kamath, Javier F. Torres-Roca, and Steven A. Eschrich  
Volume 2017, Article ID 6576840, 9 pages

**Characteristics and Validation Techniques for PCA-Based Gene-Expression Signatures**  
Anders E. Berglund, Eric A. Welsh, and Steven A. Eschrich  
Volume 2017, Article ID 2354564, 13 pages

**Module Anchored Network Inference: A Sequential Module-Based Approach to Novel Gene Network Construction from Genomic Expression Data on Human Disease Mechanism**  
Annamalai Muthiah, Susanna R. Keller, and Jae K. Lee  
Volume 2017, Article ID 8514071, 9 pages

**A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data**  
Jennifer D. Hintzsche, William A. Robinson, and Aik Choon Tan  
Volume 2016, Article ID 7983236, 16 pages

**GPA-MDS: A Visualization Approach to Investigate Genetic Architecture among Phenotypes Using GWAS Results**  
Wei Wei, Paula S. Ramos, Kelly J. Hunt, Bethany J. Wolf, Gary Hardiman, and Dongjun Chung  
Volume 2016, Article ID 6589843, 6 pages

**Embracing Integrative Multiomics Approaches**  
Daniel M. Rotroff and Alison A. Motsinger-Reif  
Volume 2016, Article ID 1715985, 5 pages

**Clinical Application of a Modular Genomics Technique in Systemic Lupus Erythematosus: Progress towards Precision Medicine**  
Eric Zollars, Sean M. Courtney, Bethany J. Wolf, Norm Allaire, Ann Ranger, Gary Hardiman, and Michelle Petri  
Volume 2016, Article ID 7862962, 7 pages
The Promise of Genomic Studies on Human Diseases: From Basic Science to Clinical Application

Lam C. Tsoi, Bethany Wolf, and Y. Ann Chen

1Department of Dermatology, University of Michigan Medical School, Ann Arbor, MI, USA
2Department of Computational Medicine & Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA
3Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA
4Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA
5Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL, USA

Correspondence should be addressed to Lam C. Tsoi; alextsoi@umich.edu

Received 12 March 2017; Accepted 12 March 2017; Published 29 March 2017

Copyright © 2017 Lam C. Tsoi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

The advances in biotechnologies and efficiency in computational resources have provided unprecedented opportunities to study and analyze the genomics of human diseases. Over the last decade, high-throughput experiments studying -omics (e.g., genetics, epigenetics, or transcriptomics) have been used to generate informative data researchers can use to test different data-driven hypotheses. A big promise of such high-dimensional -omics data is the advancement of biomedicine by effectively translating findings from basic science research into clinical application. Designing and conducting genomic experiments in biomedical research aim to enhance the diagnosis, treatment, and prevention of human diseases. Translating -omics findings into clinical practice requires a flexible framework to incorporate different -omics data types to predict clinical outcomes in an integrated fashion.

2. Data Analysis

Developing rigorous statistical approaches and implementing innovative computational tools play essential roles in translating the findings based on high-dimensional -omics data into accurate and informative medical decisions. To equip readers with updated analytical approaches, this special issue covers a wide range of analytical approaches and pipelines. J. D. Hintzsche et al. provided a comprehensive review of computational tools to analyze and interpret the whole exome sequencing (WES) data, including alignment, variant calling, and annotation approaches developed for “pre-VCF (variant calling file)” analyzes, as well as major approaches to conduct downstream analysis after VCF file has been generated: pathway analysis, somatic prediction, copy number estimation, and so forth. Robustness and the ability to replicate findings in independent datasets are also critical in analyzing high-dimensional data. For analyzing transcriptomic data, A. E. Berglund et al. proposed a principal component analysis- (PCA-) based technique to reveal gene expression signatures that are robust in replicated datasets. The method can also identify complex signatures from independent biological components. Beyond traditional data analysis, W. Wei et al. demonstrated that visualization is a key component to translate -omics data into useful information. The study utilized the GPA (genetic analysis incorporating pleiotropy and annotation) and MDS (multidimensional scaling) techniques to illustrate genetic relationships between different human traits/diseases, revealing the underlying shared genetic architecture.

3. Data Integration

Data integration is essential for robust modeling of complex or heterogeneous conditions. By integrating gene expression data with prior biological knowledge such as tissue of origin or mutation status, V. P. Kamath et al. illustrated enhanced performance on radiation sensitivity. Their results provide a proof of concept on how accounting for biological heterogeneity can lead to robust modeling of clinical response.
D. M. Rotroff and A. A. Motsinger-Reif reviewed current data integration techniques for joint-analysis of multiple -omics data and discussed future directions and challenges for applying these integrative approaches in personalized medicine. K. Raja et al. then discussed how researchers can utilize the large volume of data from the literature to develop biological inference for -omics analysis by providing an in-depth review of text-mining approaches that can be used to synthesize biomedical or clinical information and also highlighted the applications of text-mining in genomic, proteomic, and transcriptomic studies.

4. Biological and Clinical Inference

A. Muthiah et al. proposed a novel inference technique, called Module Anchored Network Inference (MANI), to reveal gene-gene relationships and provide inference on disease mechanism by using time-series gene expression data on adipocyte differentiation. Instead of utilizing all candidate genes from data, the MANI approach constructs small network modules, which is shown to outperform other in silico network inference techniques. Many human diseases are heterogeneous in nature and thus are challenging to provide accurate diagnosis and monitoring. Using systemic lupus erythematosus as a disease model, E. Zollars et al. applied a genomic technique to develop robust biomarker signature to better monitor disease activity. The approach utilized gene expression profiles to guide the classification of patients with different disease activities. In addition to genomic information, T. Nishihori and K. Shain reviewed how integrating molecular information can advance treatment of multiple myeloma in clinical setting using a risk-adapted strategy.

5. Conclusion

This special issue presents and discusses technological and methodological developments in biomedical research leading to advances in biomedicine through analysis and evaluation of -omics data. The research and review articles provide a comprehensive collection of approaches and studies for translating biological information from high-dimensional data to clinical applications. With the explosion of big data, we believe that innovative techniques, rigorous analytical approaches, and pipelines are keys to provide robust findings that can advance their clinical applications.

Acknowledgments

The guest editors would like to thank all the authors that contributed to this special issue; their scientific findings and insights have made this special issue possible.

Lam C. Tsoi
Bethany Wolf
Y. Ann Chen
In the past decade, the volume of “omics” data generated by the different high-throughput technologies has expanded exponentially. The managing, storing, and analyzing of this big data have been a great challenge for the researchers, especially when moving towards the goal of generating testable data-driven hypotheses, which has been the promise of the high-throughput experimental techniques. Different bioinformatics approaches have been developed to streamline the downstream analyzes by providing independent information to interpret and provide biological inference. Text mining (also known as literature mining) is one of the commonly used approaches for automated generation of biological knowledge from the huge number of published articles. In this review paper, we discuss the recent advancement in approaches that integrate results from omics data and information generated from text mining approaches to uncover novel biomedical information.
only over 10,000 research articles published prior to 2006 by using the same search phrase. However, the acquired data raises various significant challenges: (i) the interpretation of high-throughput results; (ii) the translation of biological data to clinical application; (iii) the data handling, storage, and sharing issues; and (iv) the reproducibility when comparing between different experiments [11, 12]. Among these, the last challenge has been a long-lasting issue, most likely due to the potential discrepancies in processing and interpreting the high-throughput data or due to “cherry-picking” approach to subjectively focus on the components that are indeed false positives. The traditional strategies to overcome these challenges are to conduct extensive literature search and seek professional opinions from domain experts to decipher the mechanism and then conduct downstream experiments to verify the findings. However, this has proven to be time consuming and subjective and has not been a common practice when researchers publish their results from high-throughput experiments. On the other hand, automated approaches have gained much interest in recent years to annotate gene functions [13], to identify biomarkers [14], and to explore genetic mutations [15]. Text mining (also known as literature mining) is a technique that has been used to retrieve and process research articles from PubMed database and can summarize biomedical information present across articles. In molecular biology, text mining is typically used to retrieve relevant documents, prioritize the documents, extract the biomedical concepts (e.g., genes, proteins, cell, tissue, and cell-type), and extract the causal relationships between concepts [16, 17]. Text mining can significantly decrease the time and effort required, compared with traditional labor-intensive approaches.

In this review, we first discuss the various omics techniques used in healthcare and summarize the recent advances in utilizing text mining approaches to facilitate the interpretation and translation of these omics data. We then focus on biomedical literature mining and clinical text mining and further describe the challenges involved in integrating the knowledge from different resources to enhance the biomedical research. Finally, we explain the recent methods to integrate omics and biomedical literature mining data in order to uncover novel biomedical information.

2. The Study of “Omics”

Traditionally, “omics” corresponds to the study of four major biomolecules: genes, proteins, transcriptomes, and metabolites [4]. Since the discovery of DNA [31], much interest has been gained towards understanding the roles of genes and proteins in cellular functions and transduction. Healthcare is considered to vary from one individual to another based on his genome, proteome, transcriptome, and metabolome. The digital revolution has paved the way for integrating patient omics data with the findings in literature for the discovery of novel biomarkers and drug targets [32–34]. Therefore, the study of omics has expanded beyond these four major omics studies, and Table 1 summarizes the various types of omics data applied to biomedical discoveries. The study of omics has introduced the realm of big data to biomedicine [35, 36]. While the first human genome project took more than a decade to complete and involved $3 billion dollars, the entire genome can be sequenced and analyzed within hours for ~$1000 now. Thus, biomedical projects are now possible to generate information at the petabyte (i.e., 1,012 bytes) scale. Nevertheless, the greatest challenge is the large-scale data analysis and its integration with clinical data available in patient electronic health records (EHR) [37].

Cloud [38] and parallel computing [39] are currently used in omics research to handle the huge volume of data. Cloud computing is described as a network of computers connected together through the Internet for effective processing. It is available remotely, through cloud computing providers (e.g., Microsoft, Google, and Amazon), and researchers have an option to make use of it at an affordable cost. Parallel computing speeds up the processing time using the same hardware and Internet setup. The combined approach of using cloud computing and parallel computing together is capable of processing omics data in a feasible time [40, 41]. Other high performance computing platforms include clusters [42], grid computing [43], and graphical processing units [44]. Processing omics data and applying bioinformatics models to the data require expertise to integrate computational, biological, mathematical, and statistical knowledge.

3. Text Mining

PubMed database is a main repository for biomedical literature and contains over 26 million articles. The number of articles being published and indexed by PubMed is increasing exponentially, and therefore text mining has become an attractive (and standard) approach in mining literature data when comparing with the traditional labor-intensive strategies. Researchers use the text mining approach to tackle information overload, both in biomedical and in general areas of big data collection, because it automates data retrieval and information extraction from the unstructured biomedical texts to reveal novel information [45, 46]. While information extraction examines the relationships between specific kinds of information contained within or between documents, information retrieval focuses on summarizing data from the larger units of documents [47]. Another automated approach to deal with unstructured data is Natural Language Processing (NLP). While text mining concentrates on solving a specific problem in a particular domain, NLP attempts to understand the text as a whole [48]. Recently, text mining and NLP have been used to address different biological questions in omics research [49].

3.1. Biomedical Literature Mining. The era of applying text mining approaches to biology and biomedical fields came into existence in 1999. It was first applied to the biomedical domain for gene expression profiling [50], as well as the extraction and visualization of protein-protein interaction [51]. It emerged as a hybrid discipline from the edges of three major fields, namely, bioinformatics, information science, and computational linguistics. Biomedical literature
<table>
<thead>
<tr>
<th>Omics</th>
<th>Study topic</th>
<th>Biomedical applications[^†]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetics/molecular genetics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genomics</td>
<td>Genes</td>
<td>Gencode, Entrez Gene</td>
</tr>
<tr>
<td>Epigenomics</td>
<td>Epigenetics modifications</td>
<td>Gene Express Omnibus</td>
</tr>
<tr>
<td>Exposomics</td>
<td>Disease-causing environmental factors</td>
<td>Comparative Toxigenomics Database</td>
</tr>
<tr>
<td>Exomics</td>
<td>Exons in a genome</td>
<td>ICE—a human splice sites database</td>
</tr>
<tr>
<td>ORFeomics</td>
<td>Open Reading Frame (ORF)</td>
<td>Human Phenotype Ontology</td>
</tr>
<tr>
<td>Phenomics</td>
<td>Phenotypes</td>
<td>PharmGKB</td>
</tr>
<tr>
<td>Pharmacogenomics</td>
<td>Impact of genes on individual's response to drugs</td>
<td>—</td>
</tr>
<tr>
<td>Pharmacogenetics</td>
<td>SNPs and their impact on pharmacodynamics and pharmacokinetics</td>
<td>—</td>
</tr>
<tr>
<td>Toxicogenomics</td>
<td>Genes response to toxic substances</td>
<td>—</td>
</tr>
<tr>
<td>Molecular biology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proteomics</td>
<td>Proteins and amino acids</td>
<td>Proteomics Identifications Database (PRIDE)</td>
</tr>
<tr>
<td>Metabolomics</td>
<td>Metabolites</td>
<td>HMDB: Human Metabolome Database</td>
</tr>
<tr>
<td>Transcriptomics</td>
<td>Transcripts (i.e., rRNA, mRNA, tRNA, and microRNA)</td>
<td>Human Transcriptome Map</td>
</tr>
<tr>
<td>Ionomics</td>
<td>Inorganic biomolecules</td>
<td>—</td>
</tr>
<tr>
<td>Kinomics</td>
<td>Protein kinases</td>
<td>KinBase database and KinWeb database</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>Genetic material from multiple organisms</td>
<td>MG-RAST</td>
</tr>
<tr>
<td>Regulomics</td>
<td>Transcription factors and other biomolecules involved in the regulation of</td>
<td>miRegulome</td>
</tr>
<tr>
<td></td>
<td>gene expression</td>
<td>—</td>
</tr>
<tr>
<td>Toponomics</td>
<td>Cell and tissue structure</td>
<td>—</td>
</tr>
<tr>
<td>Medicine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trialomics</td>
<td>Human interventional trials from clinical trials</td>
<td>—</td>
</tr>
<tr>
<td>Connectomics</td>
<td>Structural and functional connectivity in brain</td>
<td>—</td>
</tr>
<tr>
<td>Interactomics</td>
<td>Interferons</td>
<td>CREDO</td>
</tr>
</tbody>
</table>

[^†]: The list shows example applications.
mining is concerned with the identification and extraction of biomedical concepts (e.g., genes, proteins, DNA/RNA, cells, and cell types) and their functional relationships [17]. The major tasks include (i) document retrieval and prioritization (gathering and prioritizing the relevant documents); (ii) information extraction (extracting information of interest from the retrieved document); (iii) knowledge discovery (discovering new biological event or relationship among the biomedical concepts); and (iv) knowledge summarization (summarizing the knowledge available across the documents). A brief description of the biomedical literature mining tasks is listed as follows.

Biomedical Text Mining Tasks

**Document Retrieval.** The process of extracting relevant documents from a large collection is called document retrieval or information retrieval [52]. The two basic strategies applied are query-based and document-based retrieval. In query-based retrieval, documents matching with the user specified query are retrieved. In document-based retrieval, a ranked list of documents similar to a document of interest is retrieved.

**Document Prioritization.** The retrieved documents are usually prioritized to get the most relevant document. Many biomedical document retrieval systems achieve prioritization based on certain parameters including journal-related metrics (e.g., impact factor, citation count) [53] and MeSH index [54, 55] for biomedical articles. The similarity between the documents is estimated with various similarity measurements (e.g., Jaccard similarity, cosine similarity) [56].

**Information Extraction.** This task aims to extract and present the information in a structured format. Concept extraction and relation/event extraction are the two major components of information extraction [57, 58]. While concept extraction automatically identifies the biomedical concepts present in the articles, relation/event extraction is used to predict the relationship or biological event (e.g., phosphorylation) between the concepts [59, 60].

**Knowledge Discovery.** It is a nontrivial process to discover novel and potentially useful biological information from the structured text obtained from information extraction. Knowledge discovery uses techniques from a wide range of disciplines such as artificial intelligence, machine learning, pattern recognition, data mining, and statistics [61]. Both information extraction and knowledge discovery find their application in database curation [62, 63] and pathway construction [64, 65].

**Knowledge Summarization.** The purpose of knowledge summarization is to generate information for a given topic from one or multiple documents. The approach aims to reduce the source text to express the most important key points through content reduction selection and/or generalization [66]. Although knowledge summarization helps to manage the information overload, the state of the art is still open to research to develop more sophisticated approaches that increase the likelihood of identifying the information.

**Hypothesis Generation.** An important task of text mining is hypothesis generation to predict unknown biomedical facts from biomedical articles. These hypotheses are useful in designing experiments or explaining existing experimental results [67].

Conventional text mining approaches process PubMed abstracts rather than the full-text articles and fail to mine the information not in abstracts. Recently, text mining from the full-text articles is gaining more interest [59]. However, it involves many challenges: (1) the availability of full-text articles is limited (4 million full-text articles in PubMed Central versus 26 million abstracts in PubMed); (2) text mining within tables, figures, and equations is complicated; and (3) information redundancy within the articles. An automated text mining system is generally evaluated using a standard corpus (Table 2). However, the availability of standard corpora in biomedical domain is limited because its generation is expensive, time consuming, and requires domain experts. In general, a gold standard is developed within the research groups when the standard corpora are not available, but mostly not available to other researchers. The text mining systems are commonly evaluated using precision, recall, and f-score. Precision is defined as the relevance accuracy, recall is defined as the retrieval accuracy, and f-score is defined as the harmonic mean of precision and recall [56].

3.2. Clinical Text Mining. Electronic health records, discharge summaries, and clinical narratives of patients are rich in information that could be useful for improving the healthcare. In addition, the information is also available from the transcription of dictations, direct entry by clinicians/physicians, or speech recognition software. The encoding of structural information from the clinical resources is useful to clinicians and researchers. For example, automated high-throughput clinical applications can be developed to support clinicians’ information needs [68]. However, manual encoding is expensive and limited to primary and secondary diagnoses. Clinical text mining, also known as clinical NLP or Medical Language Processing (or simply MLP), is suggested as a potential technology by Institute of Medicine for mining clinical resources. The tasks described above in biomedical literature mining are applicable to clinical text mining and include additional subtasks [69]: (i) negation recognition (e.g., “patient denies on developing rashes”), (ii) temporal extraction (e.g., “small bumps noticed last year”), and (iii) patient-event relationship (e.g., “patient mother had arthritis”).

The modern healthcare relies on big data analytics for integrating, organizing, and utilizing different pharmacological or clinical information. A hybrid approach to combine patient genomic data and electronic health record information is expanding as the future vision of healthcare. The omics data has become an emerging tool for diagnosis/clinical investigations of common and rare diseases and helps in clinical decision making (i.e., selecting the best possible treatments for patients). Genome-Wide Association
<table>
<thead>
<tr>
<th>Corpus</th>
<th>Text mining evaluation task</th>
<th>Brief introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>JNLPBA (Joint Workshop on NLP in Biomedicine and Its Applications) [18]</td>
<td>Gene/protein concept extraction</td>
<td>The corpus consists of 2,000 PubMed abstracts as training data and 404 PubMed abstracts as test data.</td>
</tr>
<tr>
<td>BioCreAtiVe 2004 Task 1A dataset [19]</td>
<td>Gene/protein concept extraction</td>
<td>The corpus consists of 15,000 PubMed sentences as training data and 5,000 PubMed sentences as test data.</td>
</tr>
<tr>
<td>BioCreAtiVe 2 Gene Mention (GM) dataset [20]</td>
<td>Gene/protein concept extraction</td>
<td>The corpus consists of 15,000 PubMed sentences as training data and 5,000 PubMed sentences as test data.</td>
</tr>
<tr>
<td>AIMED [21]</td>
<td>Protein-protein interaction</td>
<td>The corpus consists of 225 PubMed abstracts that contain 1,987 sentences with 4,075 protein mentions.</td>
</tr>
<tr>
<td>HPRD50 (Human Protein Reference Database) [22]</td>
<td>Protein-protein interaction</td>
<td>The corpus consists of sentences with protein-protein interaction from 50 PubMed abstracts.</td>
</tr>
<tr>
<td>BioInfer (Bio Information Extraction Resource) [23]</td>
<td>Protein, gene, and RNA relationships</td>
<td>The corpus consists of 100 sentences annotated with concept names, relationships, and syntactic dependencies.</td>
</tr>
<tr>
<td>IEPA (Interaction Extraction Performance Assessment) [24]</td>
<td>Protein-protein interaction</td>
<td>The corpus consists of more than 200 PubMed sentences annotated with protein-protein interaction.</td>
</tr>
<tr>
<td>BioCreAtiVe 2.5 Elsevier Corpus [25]</td>
<td>Protein-protein interaction</td>
<td>The corpus consists of 61 PubMed articles as training data and 62 PubMed articles as test data.</td>
</tr>
<tr>
<td>BC4GO Corpus [26]</td>
<td>Gene ontology</td>
<td>The corpus consists of 1356 distinct GO terms from 200 PubMed articles.</td>
</tr>
<tr>
<td>GREC Corpus [27]</td>
<td>Gene regulation and gene expression events</td>
<td>The corpus consists of 240 PubMed abstracts with annotations on gene regulation and gene expression events.</td>
</tr>
<tr>
<td>GETM [28]</td>
<td>Gene expression events</td>
<td>The corpus consists of 150 PubMed abstracts with annotation for gene expression events.</td>
</tr>
<tr>
<td>AnEM [29]</td>
<td>Tissue, cell, developing anatomical structure, cellular component</td>
<td>The corpus consists of 500 PubMed sentences with annotations on variety of biomedical concepts.</td>
</tr>
<tr>
<td>CellFinder Corpus [30]</td>
<td>Anatomical parts, cell lines, cell types, species, and cell components</td>
<td>The corpus consists of annotations from 10 full-text PubMed articles.</td>
</tr>
</tbody>
</table>
study (GWAS), also known as Whole Genome Association Study (WGAS), is a relatively new approach for identifying genes (i.e., loci associated with human traits) through rapid scanning of markers across whole DNA or genome [70]. GWAS has been applied also to cancer research for drug repositioning [71], prioritizing susceptible genes in Crohn's disease [72], and analyzing the human variants in the area of precision medicine [73]. As an example, the Michigan Genomics Initiatives (MGI) at the University of Michigan has developed an institutional based DNA and genetics repository combined with patient phenotype. The project aims to bring awareness to each patient/participant about the disease development and response to treatments for better health and wellness. The current studies at MGI include analgesics outcome study (AOS), understanding opioid use in chronic pain patients, a pivotal study on high-frequency nerve block for postamputation pain, Michigan body map (MBM), and positive piggy bag (https://www.michiganengenomics.org).

Clinical text mining faces the following specific challenges: (1) access to patient EHR requires permission from Institutional Review Board (IRB); (2) personal details of the patients should be deidentified; (3) mining approaches developed for mining gene/protein interactions (e.g., gene-chemical, gene-disease) [79, 80] to support the complexity of available information and coverage of varieties of biomedical subdomains [16]. Text mining approaches do not replace the manual curation of biomedical information but support speeding up the process by several-fold [75, 76]. In this section we describe the various text mining approaches developed for mining omics related information.

4. Role of Text Mining in Omics Study

Relationship between concepts of the same kind (e.g., gene-gene) or different kind (e.g., gene-disease) is commonly known as “event” [74]. The events are useful to identify many clinical facts such as disease onset and response to drug treatment. Overwhelming of biomedical articles from omics research has accumulated abundance of information and requires advanced event extraction systems to support the complexity of available information and coverage of varieties of biomedical subdomains [16]. Text mining approaches do not replace the manual curation of biomedical information but support speeding up the process by several-fold [75, 76]. In this section we describe the various text mining approaches developed for mining omics related information.

4.1. Genomics and Text Mining. In the current era of genomics, text mining plays an important role in mining gene-gene interactions [77, 78] and other gene involved interactions (e.g., gene-chemical, gene-disease) [79, 80] to support integrative analysis of gene expression [81, 82], pathway construction [83, 84], ontology development [85], and database annotation [62, 86, 87].

Genes encode proteins and proteins enroll in various biological functions by interacting with other proteins. This encoding process is defined in two steps: transcription (i.e., DNA to RNA) and translation (RNA to protein). Many cellular processes are regulated by microRNA through mRNA degradation and suppression of gene expression such that the protein synthesis is interrupted. This is the fundamental of genomics. In genomics, gene function is assessed from the involvement of genes/proteins in biochemical pathways. The functional genomics is a revolutionary area in text mining where the gene/protein mentions in the biomedical articles and their relationship are considered to be important. Furthermore, gene and protein names are highly complex and text mining has contributed to their recognition in the unstructured text [57, 58].

Different text mining implementations for exploring the finding of genome research have been developed in the past decade. miRTex is a text mining system developed for mining experimentally validated microRNA gene targets from PubMed articles. The system has been successfully implemented to identify the Triple Negative Breast Cancer related genes that are regulated by microRNAs [81]. More sophisticated approaches integrate gene expressions from microarray experiments, biomedical data extracted by text mining, and gene interaction data to predict gene-based drug indications [82]. A similar approach [87] attempts to support manual curation of links between biological databases such as Gene Expression Omnibus (GEO) and PubMed database. Another approach [88] combines text mining data with microarray data for discovering disease-gene association by using unsupervised clustering. The gene-drug interaction information extracted by text mining is used to predict the drug-drug interaction [89]. Above all, the researchers have attempted to use text mining for annotating genome function with gene ontology [90]. Thus, text mining and genomics together uncover much biomedical information that was previously unknown.

4.2. Proteomics and Text Mining. Protein-protein interaction is important to explore the mechanism involved in biological processes and onset of diseases [91]. Intact [92], BIND [93], MIND [94], and DIP [95] are the major databases available for protein-protein interaction. These databases are manually curated by the domain experts, but a larger portion of information is still available only in the biomedical literature. Text mining provides a bridge to cover the gap existing between the manual curation and information hidden in the literature. The approaches to extract protein-protein interaction range from simple rule-based systems and cooccurrence systems to more sophisticated NLP methods [60] and machine learning systems [96]. Apart from protein-protein interaction extraction systems, text mining also provides automated approaches for extracting posttranslational modification of proteins such as protein phosphorylation [59].

4.3. Transcriptomics, Metabolomics, and Text Mining. Text mining approaches for transcriptomics and metabolomics are limited. One major fact is that these two areas of genomics are comparatively new when compared to genomics and proteomics. A recent study compares the metagenome characteristics of healthy individuals with autism patients to analyze the enzymes involved [97]. The computational approach uses text mining for genomics and metabolomics information.
extraction. A web-based tool called 3Omics is available for integrating, comparing, analyzing, and visualizing data from transcriptomics, metabolomics, and proteomics [98]. Another tool called Babelomics integrates transcriptomics, proteomics, and genomics data to uncover the underlying function profiles [99]. Thus, a wide variety of hidden biomedical information within the omics data are extracted and predicted through text mining.

5. Conclusion

In this review, we summarized the current state of the art in omics research and contribution of text mining approaches to uncover the omics related biomedical information hidden within the published articles. We discussed the core concepts of omics and the challenges involved in storing and analyzing the huge volume of omics data generated from high-throughput experiments. We also highlighted the use of computer techniques such as parallel processing and cloud computing to manage omics data and elaborated on text mining approaches for biomedical literature and clinical text with emphasis on omics. While the omics approach is emerging to be commonly used practice for basic science or clinical diagnosis technique, it is imminent to note that data interpretation and translation is the bottleneck. The advances in text mining can be useful to resolve the challenges with the omics data and further support in novel biomedical discoveries.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors acknowledge the support from the Undergraduate Research Opportunity Program (UROP) from the University of Michigan, the Dermatology Foundation, the Arthritis National Research Foundation, and the National Psoriasis Foundation.

References

A. Tebani, C. Afonso, S. Marret, and S. Bekri, “Omics-based


J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H.

T. Ohta, S. Pyysalo, J. Tsujii, and S. Ananiadou, “Open-

M. Gerner, G. Nenadic, and C. M. Bergman, “An exploration

P. Thompson, S. A. Iqbal, J. McNaught, and S. Ananiadou,

F. Leitner, M. Krallinger, G. Cesareni, and A. Valencia, “The

S. Pyysalo, F. Ginter, J. Heimonen et al., “BioInfer: a corpus


generation sequencing,” *BMC Genomics*, vol. 14, no. 1, article no. 789, 2013.


Integrating Biological Covariates into Gene Expression-Based Predictors of Radiation Sensitivity

Vidya P. Kamath,1 Javier F. Torres-Roca,2 and Steven A. Eschrich1

1Department of Biostatistics & Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA
2Department of Radiation Oncology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA

Correspondence should be addressed to Steven A. Eschrich; steven.eschrich@moffitt.org

Received 30 September 2016; Revised 4 January 2017; Accepted 11 January 2017; Published 8 February 2017

Academic Editor: Bethany Wolf

Copyright © 2017 Vidya P. Kamath et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The use of gene expression-based classifiers has resulted in a number of promising potential signatures of patient diagnosis, prognosis, and response to therapy. However, these approaches have also created difficulties in trying to use gene expression alone to predict a complex trait. A practical approach to this problem is to integrate existing biological knowledge with gene expression to build a composite predictor. We studied the problem of predicting radiation sensitivity within human cancer cell lines from gene expression. First, we present evidence for the need to integrate known biological conditions (tissue of origin, RAS, and p53 mutational status) into a gene expression prediction problem involving radiation sensitivity. Next, we demonstrate using linear regression, a technique for incorporating this knowledge. The resulting correlations between gene expression and radiation sensitivity improved through the use of this technique (best-fit adjusted $R^2$ increased from 0.3 to 0.84). Overfitting of data was examined through the use of simulation. The results reinforce the concept that radiation sensitivity is not driven solely by gene expression, but rather by a combination of distinct parameters. We show that accounting for biological heterogeneity significantly improves the ability of the model to identify genes that are associated with radiosensitivity.

1. Introduction

One of the goals of developing biomarkers is for use in patient selection, diagnosis, and management of cancer treatment [1–3]. An important aspect in management of cancer treatment is to understand how a patient will respond to a specific treatment such as radiation therapy. Designing the radiation therapy to maximize cancer cell death is beneficial, and predicting such a response of the cells to radiation therapy is important for effective patient management. Genes such as RAS [4, 5] and p53 [6] have been known to influence the response of tumor cells to radiation treatment. For example, RAS has been implicated as a central regulator of radioresistance. Similarly, presence of a mutant p53 gene is used as an indicator for uncontrolled proliferation of cells, while a wild-type p53 gene is known to be a tumor suppressor. In addition tissue of origin has been associated with radiosensitivity. For example, the SF2 (survival fraction of cells after 2 Gy of radiation) of melanoma and glioma cell lines has been shown to be higher (radioresistant) than lymphoma and myeloma cell lines [7–9].

The process of developing the systems-based model of radiosensitivity followed a stepwise strategy. The first step was to develop a radiosensitivity classifier to predict cellular radiosensitivity based on gene expression profiles [10]. We developed a multivariable linear regression model that correlated gene expression to radiosensitivity as determined by SF2, in a 35-cell line database. We used a leave-one-out cross-validation approach, where the classifier was developed using 34 of the 35 cell lines as a training set, leaving one cell line as a test set. The basal gene expression profiles and the radiation sensitivity of all 34 cell lines in the training set were used to identify genes that were correlated with radiosensitivity. This was performed using SAM analysis (Significant Analysis of Microarrays) [11] with a false discovery rate of 5%. Genes selected by SAM were then combined as radiosensitivity predictors during the construction of the classifier. A multivariable linear regression model was created...
using these probesets to predict the SF2 of the test sample and was shown to achieve a statistically significant (p = 0.002) predictive accuracy of 62%, within a continuous classification problem. The classifier predicts an actual SF2 value (range: 0.01–1.0) rather than a binary phenotype (radioresistant versus radiosensitive). Importantly, we biologically validated the model by demonstrating that three of the genes selected by the algorithm (rbap48, rgs-19, and top-1) were mechanistically involved in radiation response. Thus, we demonstrated that cellular radiosensitivity is predictable based on gene expression but more importantly we validated this approach as a strategy for the discovery of novel radiosensitivity biomarkers.

Although we had developed a successful mathematical model correlating gene expression and radiosensitivity, we reasoned the model had a number of problems that if overcome would significantly improve its ability to impact the field of radiation biology. First, expansion of the cell line dataset from 35 samples should provide more reliable correlations. Second, there were few genes consistently selected by the classifier. A larger pool of genes would be desirable, as it would allow us to identify the biological networks that regulate cellular radiosensitivity. Third, gene expression was the only variable considered in the model, while there are several biologic factors besides gene expression that are known to influence radiosensitivity. Therefore we focused on strategies aimed at increasing the pool of candidate genes and incorporating biologic variables into the algorithm. One of the advantages of developing the classifier in the NCI-60 is that these cell lines are molecularly well characterized, thus allowing the inclusion of important biological variables into the process. We chose four variables that have been previously correlated to radiation sensitivity: gene expression [10], tissue type [8, 12], RAS mutation status [13–18], and p53 mutation status [19–21]. In addition we expanded the cell line dataset from 35 to 48 cell lines.

2. Material and Methods

2.1. Microarrays. Gene expression profiles were from Affymetrix HU6800 chips (7,129 genes) from a previously published study [22]. These are publicly available as supplemental data to the published study. The gene expression data had been previously preprocessed using the Affymetrix MAS 5.0 algorithm in average difference units. Negative expression values were set to zero and the chips were normalized to the same mean intensity. Specific cell lines used are listed in Supplemental Table 1 in Supplementary Material available online at https://doi.org/10.1155/2017/6576840.

2.2. Radiation Survival Assays (SF2). The SF2 of cell lines used in model development were previously reported [10, 23]. SF2 values are included in Supplemental Table 1.

2.3. Permutation Analysis. Predictions were randomly permuted among cell lines 10,000 times and accuracies greater than or equal to the threshold were counted to calculate a p value for significance relative to chance.

2.4. Gene Expression Model. Gene expression and radiation sensitivity were described through a linear relationship as described in (1). In this equation, SF2n represents the radiation sensitivity (as measured by SF2) for cell line n in the dataset. kn represents a model coefficient, computed during the training process, and yni represents the gene expression value for the ith probeset for cell line n. The least-squares fit of the individual linear models was compared when selecting probesets of interest for modeling radiosensitivity.

Gene Expression-Only Model

\[
SF2_n = k_0 + k_1(y_{ni}) .
\]  

(1)

2.5. Inclusion of Biological Covariates in Model Development. We hypothesized that incorporating biological covariates into the gene selection process would improve the ability of the algorithm to identify radiosensitivity biomarkers. To integrate biological covariates into model development we constructed individual gene-based models using two different equations to relate gene expression and the biological parameters to radiosensitivity (SF2). Specific biological parameters are tissue of origin (TO), RAS mutation status (RAS), and p53 mutation status (p53).

Additive Model

\[
SF2 = k_0 + k_1(TO_n) + k_2(TO_n) + k_3(RAS_n) + k_4(p53_n).
\]  

(2)

Interactive Model

\[
SF2_n = k_0 + k_1(y_{ni}) + k_2(TO_n) + k_3(RAS_n) + k_4(p53_n) + k_5(y_{ni})(TO_n) + k_6(y_{ni})(RAS_n) + k_7(y_{ni})(p53_n) + k_8(y_{ni})(RAS_n)(p53_n) + k_9(y_{ni})(TO_n)(RAS_n) + k_{10}(TO_n)(p53_n) + \cdots
\]  

(3)

In (2) and (3), the cell line radiosensitivity (SF2n) was modeled as a function of gene expression (y) and biological variables (TO, RAS, and p53). Specifically, SF2n represents the radiosensitivity of cell line n and yni represents the gene expression value of an individual probeset (i) for the nth cell line in the dataset. A total of 9 different TO values were present in the 48 cell line database. RASn and p53n were binary variables (wild-type/mutated) for the nth cell line. Thus, the additive model considered a total of 13 terms (an intercept, gene expression, 9 TO, RAS, and p53). The more complex interactive model initially considered all possible
2.6. Random Variables. Random variables for exploring the effect of RAS and p53 mutation status were created and uniformly distributed into two states (one each for the mutated and wild-type status). The frequencies of these states were similar to the true distributions in the data. Similarly, a random variable was defined for TO, with each sample being assigned a tissue type at random. This new dataset with randomly assigned biological parameters was used to test whether the improvement in linear fit achieved by both the additive and interactive model was due to the integration of biological variables or due to chance.

Table 1: Terms used in linear modeling. The term (y) represents gene expression. The operator × represents an interaction term between two or more variables.

<table>
<thead>
<tr>
<th>Terms</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>y × tissueTypeBREAST</td>
</tr>
<tr>
<td>y (gene expression)</td>
<td>y × tissueTypeCNS</td>
</tr>
<tr>
<td>tissueTypeBREAST</td>
<td>y × tissueTypeCOLON</td>
</tr>
<tr>
<td>tissueTypeCNS</td>
<td>y × tissueTypeLEUK</td>
</tr>
<tr>
<td>tissueTypeCOLON</td>
<td>y × tissueTypeMELAN</td>
</tr>
<tr>
<td>tissueTypeLEUK</td>
<td>y × tissueTypeNSCLC</td>
</tr>
<tr>
<td>tissueTypeMELAN</td>
<td>y × tissueTypeOVAR</td>
</tr>
<tr>
<td>tissueTypeNSCLC</td>
<td>y × tissueTypePROSTATE</td>
</tr>
<tr>
<td>tissueTypeOVAR</td>
<td>y × RASmut</td>
</tr>
<tr>
<td>RA5mut</td>
<td>tissueTypeBREAST × RASmut</td>
</tr>
<tr>
<td>p53mut</td>
<td>tissueTypeCOLON × RASmut</td>
</tr>
<tr>
<td></td>
<td>tissueTypeMELAN × RASmut</td>
</tr>
<tr>
<td></td>
<td>tissueTypeNSCLC × RASmut</td>
</tr>
<tr>
<td></td>
<td>tissueTypeOVAR × RASmut</td>
</tr>
<tr>
<td></td>
<td>y × tissueTypeBREAST × RASmut</td>
</tr>
<tr>
<td></td>
<td>y × tissueTypeCOLON × RASmut</td>
</tr>
</tbody>
</table>

3. Results

3.1. Expansion of Cell Line Dataset Lowers Classification Accuracy. As described above we previously developed a gene expression radiosensitivity classifier [10] as a continuous prediction rather than a binary classification problem (i.e., radiosensitive versus radioresistant). During development of the model we had observed that increasing the number of samples increased the classifier accuracy (data not shown). Thus we hypothesized that increasing the cell line dataset to 48 cell lines would result in a more accurate model. Surprisingly, the classifier technique was not as accurate when the cell line population was increased to 48 (compared to 35) cell lines. The best linear regression-based classifier using the 48 cell lines correctly classified 26/48 samples (54%) (Figure 1(a)) compared to 25/35 (71%) for the best classifier in the 35-cell line dataset. We explored the use of alternate normalization (Figure 1(b)); however the maximum accuracy was 28/48 or 58%. Additionally, we looked at alternate predictors (Figure 1(c)) but the decreased accuracy in the 48-cell line dataset was consistent. Although the results were still statistically significant in that the classifier in the 48 cell line dataset performed better than chance (p = 0.0094), we were interested in understanding the reason for the decreased accuracy.

3.2. Understanding the Influence of Confounding Factors. The decrease in classification accuracy suggested that the linear regression model based only on gene expression data did not fully represent the classification problem. We hypothesized that accounting for the biological diversity of cell lines in the database would be of importance. Several biological variables available for the NCI-60 cell lines include tissue of origin (TO), RAS mutational status (wt/mut) (RAS), and p53 mutational status (p53). These variables have been implicated in the biological regulation of radiation sensitivity [13, 24]. Among the 48 cell lines, the RAS-mutated cell lines represent only 31% (15/48) of cell lines whereas they represented 40% (14/35) in the 35-cell line database (Figure 2(a)). The p53 mutation status was also different between the two groups; 26 cell lines were p53 mutants in the 35 cell lines; however only 5 additional mutants were added, changing the proportions from 74% down to 65% of the cell line population (Figure 2(b)). Tissue of origin was similar in proportions in the two groups (Figure 2(c)). Since only one additional RAS-mutated cell line was added when increasing the dataset to 48 we first focused on determining if RAS mutation status impacted the gene selection process.

The oncogenic protein RAS has been proposed to mediate a central mechanism in radiation resistance [16]. We tested whether the presence of a RAS mutation, which usually affords a chronically active RAS protein, was an important source of variability within the dataset. This was done by determining whether the genes selected by the 35 cell line classifier were dependent or independent of RAS status. We stratified the original 35 cell lines by RAS status and performed the gene selection step (correlation of gene expression and SF2) in each group of cell lines. The three genes (rbap48, rgs-19, and r5pia) selected by the original classifier (without
RAS stratification) were previously shown to be highly useful in predicting radiosensitivity. These genes were highly ranked among the RAS-mutated cell lines but not in the wild-type lines, suggesting that the RAS-mutated cell lines were driving the classification process. RbAp48, rgs-19, and r5pia were ranked 19th, 46th, and 262nd out of 7,129 probesets by R^2 values from the RAS-mutated cell lines. In wild-type cell lines, these same genes are ranked 743rd, 758th, and 397th, respectively. Interestingly, these three genes ranked in the top 10 genes when all cell lines were considered together (5th, 1st, and 9th) (Table 2). These results suggest that the biological diversity of cell lines studies (e.g., RAS-mutated and RAS wt) can significantly impact the evaluation of genes with respect to outcomes. In particular, two diverse biological types mixed in different proportions can lead to highly variable ranking as demonstrated by our 35-cell line experiment.
### 3.3. Integrating Biological Covariates

As a result of the analysis of confounding factors, three variables (TO, RAS, and p53) were integrated in the gene expression analysis using two approaches: an additive model and an interaction-based linear model. The gene selection process was repeated using these approaches on the 48 cell lines. RAS and p53 status indicators were binary variables that indicate wild-type (wt) or mutational (mut) status of the gene for a cell line. The indicator for tissue of origin (TO) has 9 levels, one for each type of tissue from which the tumor cell line originated [22]. The analysis was performed for each probeset and the model fit parameter adjusted-$R^2$ (Adj-$R^2$) was used to determine if the model improved by inclusion of the covariates. The adjusted-$R^2$ was used instead of $R^2$ in these experiments to adjust for addition of regressors in the equations.

Figure 3 shows a box plot summarizing the Adj-$R^2$ values from all probeset models individually when correlated with radiation response (SF2) in the 48-cell line database. In the gene expression-only model, fewer probesets had a model fit better than 0.2 (<30 of the 7129 probesets), with the best fit being just above 0.3. The average fit for the additive model was 0.28 with a maximum fit of 0.48. With the interactive model, the average fit improved to 0.6 with a maximum value of 0.84. Thus the integration of biological variables, including all interactions, improved the modeling fit considerably.

### 3.4. Verification of Model Fit

The improvement in Adj-$R^2$ of a linear model could be attributed simply to the addition of more variables in the model (e.g., overfitting) [25]. We compared the fit of the expanded linear models to the fit
Table 3: Change in Adj-$R^2$ value obtained by adding terms and complexity to the linear model. Results obtained with clinical indicators TO, RAS, and p53 are compared to Adj-$R^2$ values obtained using random variable for each indicator.

<table>
<thead>
<tr>
<th>Model terms</th>
<th>Model comparison</th>
<th>Mean $\Delta R^2$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clinical indicators</td>
<td>Random variables</td>
</tr>
<tr>
<td>GeneEx:TO</td>
<td>GenEx only versus additive</td>
<td>0.254</td>
</tr>
<tr>
<td></td>
<td>Additive versus interaction</td>
<td>0.134</td>
</tr>
<tr>
<td>GeneEx:RAS</td>
<td>GenEx only versus additive</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>Additive versus interaction</td>
<td>0.030</td>
</tr>
<tr>
<td>GeneEx:p53</td>
<td>GenEx only versus additive</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>Additive versus interaction</td>
<td>0.016</td>
</tr>
<tr>
<td>GeneEx:TO:RAS</td>
<td>Basic versus additive</td>
<td>0.256</td>
</tr>
<tr>
<td></td>
<td>Additive versus interaction</td>
<td>0.272</td>
</tr>
<tr>
<td>GeneEx:TO:p53</td>
<td>Basic versus additive</td>
<td>0.262</td>
</tr>
<tr>
<td></td>
<td>Additive versus interaction</td>
<td>0.198</td>
</tr>
<tr>
<td>GeneEx:RAS:p53</td>
<td>Basic versus additive</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>Additive versus interaction</td>
<td>0.042</td>
</tr>
<tr>
<td>GeneEx:TO:RAS:p53</td>
<td>Basic versus additive</td>
<td>0.265</td>
</tr>
<tr>
<td></td>
<td>Additive versus interaction</td>
<td>0.317</td>
</tr>
</tbody>
</table>

The difference between including biological variables and random variables in the interaction-based models is more significant. For example, the change in $R^2$ for the additive model using RAS, TO, and gene expression was similar to that of random variables; however in the interaction model, the correlation improves by 0.272 whereas the interaction of random variables (for TO and RAS) drops by 0.213 ($\Delta R^2 = -0.213$). When including all three terms in the interaction models, the Adj-$R^2$ improves by 0.317 but the random variables cause a drop in correlation ($\Delta R^2 = -0.103$).

Figure 4 summarizes the trend that when two or more biological variables are considered, this results in better linear models than expected from randomly generated variables. The interaction of random variables with gene expression data alone provides a marginal improvement in the fit; however, when two or more random variables interact, the lack of information in each variable translates into poorer fit of the linear model to the radiation sensitivity outcome. In contrast, the interaction of the biological variables adds more information to the linear model, as shown by the improvement in Adj-$R^2$ values in Table 3 and Figure 4.

4. Discussion

The central aim of our research efforts is the development of a systems biology-based understanding of the biological
In this study we show that accounting for biological confounders within a linear regression model of radiation sensitivity significantly improves the ability of the algorithm to fit gene expression and radiosensitivity, which resulted in a better ability to identify significant genes. We showed that simply adding the biological variables did not improve the fit more than what was expected from chance but when a more complex interaction-based model was utilized, its performance was superior to chance.

Although unexpected, the previously developed predictor of radiation sensitivity performed much worse when the cell line set was expanded to 48 cell lines. The underlying cause of this difference was determined to be from shifting proportions of cell lines with key biological characteristics that have been previously implicated in modulating radiation response. Specifically, we show that RAS-mutated cell lines had a large impact in gene selection in the 35-cell line dataset. Three of the top genes selected by the 35-cell line classifier were highly ranked by RAS-mutated cell lines but not by RAS wt cell lines. However when the model was expanded to 48 cell lines, the impact of these cell lines was diluted with the addition of predominantly RAS wt cell lines. Once these factors were accounted for within the modeling process, genes were identified as related to radiation sensitivity most significantly through interactions with these biological characteristics. This work led to the development of a systems-based predictive model of tumor intrinsic radiosensitivity that was validated in three independent clinical cohorts of patients treated with chemoradiotherapy [28]. However, a key insight into this process was the identification and incorporation of confounders.

The strategy presented here may have applications in the development of clinical predictive/prognostic models. For example, we have already shown that this process led to development of a predictive model of intrinsic radiosensitivity that has been clinically validated [23, 28–35]. However, clinical cohorts very often present similar diversity as that represented in the cell line database utilized and identifying key biological covariates and a mathematical approach to account for them might significantly enhance our ability to develop predictive models with clinical utility.

The inclusion of biological variables significantly improved the ability of most genes to describe the relationship between gene expression and radiation response in a linear regression model. However, the inclusion of additional parameters and their interactions within the same equation almost certainly leads at least in some instances to overfitting. It is important to note that the selection process of genes for further validation (e.g., by choosing the best-fit genes) does not require overfitting to be completely removed. Rather, it is expected that overfitting will uniformly increase the fit of genes with radiation sensitivity. In addition, the behavior of the random variables in the interaction models clearly indicates that the biological variables do provide meaningful information, and rather than causing overfitting of the model to the data, the biological variables can be used to create a better model for gene selection.

Finally, an improvement in linear fit should be similarly obtained when adding a randomly generated variable into the model instead of a variable that carries biological significance. However, it is intriguing that not all variables considered had similar impact in improving the model, as might be expected due to chance. For example RAS was significantly more important than p53 in improving the model. This observation suggests that at least part of the improvement obtained by the expanded linear models is due to a better representation of biology.

5. Conclusion

In conclusion, we demonstrate that incorporating biological covariates into a gene expression model of tumor intrinsic radiosensitivity can improve the modeling process. Accounting for biological heterogeneity can identify genes that are associated with radiosensitivity, which in turn led to the development of a successful model of clinical response to radiotherapy [23, 28–35].
Competing Interests

Javier F. Torres-Roca and Steven A. Eschrich hold patents and are cofounders of Cvergenx, Inc.

References


Research Article

Characteristics and Validation Techniques for PCA-Based Gene-Expression Signatures

Anders E. Berglund, Eric A. Welsh, and Steven A. Eschrich

Department of Biostatistics and Bioinformatics, Division of Population Sciences, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA

Correspondence should be addressed to Anders E. Berglund; anders.berglund@moffitt.org

Received 27 September 2016; Revised 15 December 2016; Accepted 4 January 2017; Published 6 February 2017

Academic Editor: Bethany Wolf

Copyright © 2017 Anders E. Berglund et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background. Many gene-expression signatures exist for describing the biological state of profiled tumors. Principal Component Analysis (PCA) can be used to summarize a gene signature into a single score. Our hypothesis is that gene signatures can be validated when applied to new datasets, using inherent properties of PCA.

Results. This validation is based on four key concepts. Coherence: elements of a gene signature should be correlated beyond chance. Uniqueness: the general direction of the data being examined can drive most of the observed signal. Robustness: if a gene signature is designed to measure a single biological effect, then this signal should be sufficiently strong and distinct compared to other signals within the signature. Transferability: the derived PCA gene signature score should describe the same biology in the target dataset as it does in the training dataset.

Conclusions. The proposed validation procedure ensures that PCA-based gene signatures perform as expected when applied to datasets other than those that the signatures were trained upon. Complex signatures, describing multiple independent biological components, are also easily identified.

1. Introduction

The use of gene signatures and Principal Component Analysis [1] (PCA) is a popular combination, but a recent publication has clearly shown drawbacks with this combination [2]. Gene signatures are used to represent a biological event and have the potential to describe complex biology better and more robustly than a single gene. There exists a large amount of literature on how to properly analyze microarray data and derive signatures [3–7], validate biomarkers [3, 8, 9], and, in particular, validate prognostic models [10–13], all in response to the poor reproducibility rate in publications [14–18]. A recent report by the Institute of Medicine [19] summarizes many of these issues of signature reproducibility and validation. We will, in this article, focus on how to quantitate the validity of applying PCA-based gene signatures to new datasets. PCA is a technique that reduces a high-dimensional dataset to a low-dimensional dataset while retaining most of the variation in the data. These new variables are referred to as scores, $t$, and the importance (weighting) of the original variables given in the loadings, $p$. For a more in-depth discussion of PCA, we refer to a recent tutorial by Bro and Smilde [20]. PCA models can describe unintended biology when there is large variation due to sources other than the biological process of interest, which can cause random signatures to be significantly associated with outcome [2].

Gene based signatures can be derived using several different techniques. One technique is to include all genes known to be involved in a specific pathway or process, such as a signaling pathway, and treat the signature as a representation of pathway activation. Gene signatures can also be derived from cell line experiments, where a specific biological event is being modulated, or by comparing different known mutation types within the cell lines. In this case, the signature is usually derived from a list of statistically significant differentially expressed genes.

In cancer research, these gene signatures can be used, for example, to predict tumor chemotherapy resistance, aggressiveness, and several other types of clinically relevant scores. These scores are then commonly used for survival
analysis, such as Kaplan–Meier plots and log-rank tests. In order to associate these clinical metrics to a gene signature, a single score value is commonly calculated from the individual expression levels of all genes within the signature. One commonly used technique is PCA, where the score vector for the first component is used to represent the gene signature. PCA is a well-established technique for data analysis and has been widely used in many areas. The resulting score can be seen as a weighted average, where each gene is weighted by its importance within the first principal component. There are several advantages to using PCA. For example, not all genes are weighted equally—more important (statistically and, we assume, biologically) genes are weighted higher. It is also robust to noise and can handle both up- and downregulated genes equally well. There are also some pitfalls, due to the inherent properties of PCA, that require special attention. One such pitfall is that if the gene signature is biologically complex, describing more than one biological event, the score from the PCA model may only describe one of the biological events. The PCA model does not necessarily describe all the biological events in the first principal component. Another, probably more well known, issue with PCA models is sign-flipping. The sign of the score value for a sample may change, or flip, depending on the software used and/or small changes in the data. This does not change the interpretation of the PCA model, but it may lead to trouble when comparing different PCA models derived from different datasets and/or software, where the signs may be reversed. The PCA model can easily be flipped back by multiplying both the scores and loadings with $-1$, a 180-degree rotation.

These issues clearly demonstrate the need to define the ideal characteristics of a PCA-based gene signature and measures of how “well behaved” a signature is when applied to a dataset. Currently, one common way to measure the performance of a signature within cancer tumor datasets is by survival analysis. Even if this is the ultimate goal of a signature, it may be misleading, as was recently shown by Venet et al. [2]. They showed that many random gene-set PCA models were as good as literature-derived and experimentally derived signatures in predicting survival. They ascribed this effect to a proliferation-signature bias present in many tumor datasets. Due to the large number of genes affected by proliferation pathways in tumors, together with large differences in proliferation status between samples, the first principle component of a PCA analysis, PC1, often captures proliferation-related effects in addition to any effects related to the signature of interest. This can result in false-correlation with survival, where the correlation comes more from proliferation-bias effects, rather than from the signature of interest. Due to these issues, it has been shown that gene signatures can be unstable [21] and that single genes can be as good as a multigene signature [22], leading to a recent review on the value of gene-expression signatures by Chibon [14]. This manuscript describes the key aspects of a PCA-based signature, along with a set of measures and figures that will describe how suitable a gene signature is when applied to a given dataset. This includes measures on how robust it is, if the signature is too complex to work well with a PCA model, if the signature differs from the general direction in the dataset, and, most importantly, if the signature describes the same biology that it was intended to. We define a gene signature as a list of genes with corresponding direction and if relevant, magnitude, which are used to describe a biological signal, such as tumor aggressiveness, distant metastasis, survival, or gender.

2. Material and Methods

Principal Component Analysis was performed using MATLAB. Gene-expression data consisted of log2 intensities. Mean-centering and unit variance scaling were applied to expression values prior to computing PCA models.

2.1. Generation of Randomized Gene Signatures. Results for the gene signature PCA model are compared to thousands of PCA models, based on randomly selected gene signatures. This approach has many benefits, including (1) the performance statistics of PCA, including explained variance, are dependent on many factors and are, thus, not comparable across different datasets and gene signatures. (2) Many of the datasets used are biased in one or several ways, as clearly shown by Venet et al. [2]. Using randomized gene signatures as comparisons removes this bias. (3) It allows for a clear measurement of how much “better” the true gene signature is than randomly generated gene signatures.

Randomized gene signatures were generated by randomly selecting an equal number of genes as the true signature and performing a PCA model for this random gene signature. This is repeated many times to get a distribution of expected values. In this work, we have chosen to use 10,000 randomized gene-sets.

2.2. Signatures. We have, in this study, used two signatures, along with some modifications of these signatures, in order to exemplify our results. The two signatures, gender and tumor versus normal, were chosen due to their strong biological signal, common use cases, and potential problems. The tumor versus normal signature largely captures proliferation/cell cycle biology, which is often a dominant feature of tumor biology [2], and gender may be used in quality control analysis, where the gender derived from molecular data is compared to the clinically known gender to identify potential sample mismatches. Each signature was also chosen to highlight a different potential pitfall of PCA analysis: nonuniqueness of signature (tumor versus normal) and absence of meaningful signal when the biology of interest does not vary (gender).

The first signature is a gender signature (Gender-29), consisting of 29 probesets representing 20 unique genes, 19 of which are present in the TCGA RNAseq dataset. To describe the directionality of these genes, we used either $-1$, for male-specific genes, or 1, for female-specific genes. The second signature was derived from a publically available Tumor/Normal dataset (GSE18842). An all-probeset PCA model of GSE18842 shows a clear separation between tumor and adjacent normal samples in the first principal component (data not shown). The Tumor/Normal (TvsN-100) gene signature was derived by taking the top 100 probesets with
the largest absolute loading values for the first principal component. Additionally, we only selected probesets that were also present on the Affymetrix U133A chip. The individual loading value from the first principal component was used to both describe the directionality and the magnitude for each gene. Since proliferation is the major biological process that is different between tumors and adjacent normals, this signature will be related to proliferation. When applied to the TCGA RNAseq dataset, 84 unique genes were used for this signature. The TvsN-100 was also used to generate modified signatures by gradually adding probesets not related to the separation of Tumor/Normal samples. This was done to simulate the addition of random noise to the signature. These are named TvsN-50/50, TvsN-25/75, and TvsN-10/90, respectively, corresponding to the percentages of TvsN-100/unrelated probesets. The last signature is a mix of the Gender-29 and the TvsN-100 signature (Mix-29/29). It was generated by merging all 29 probesets from the Gender-29 signature with the top 29 probesets for the TvsN-100 signature. For this signature, only the directionality was used. To further investigate the effect of mixing two signatures, the number of TvsN-100 probesets was gradually decreased, creating three additional datasets, Mix-29/24, Mix-29/19, and Mix-29/14. These mixed signatures will exemplify a complex signature describing more than one type of biology.

2.3. Datasets. Three publically available datasets were used in this study. These were chosen for several reasons: DC (lung tumors, both genders), GSE2034/Breast (same tumor type as data used in Venet et al. [2], single gender), and PRAD (prostate adenocarcinomas, single gender). The rationale for selecting these three datasets is the following: (1) the TvsN-100 signature is expected to perform well on all three datasets, (2) gender signature is expected to perform well on the DC dataset but not on Breast and PRAD datasets, (3) GSE2034/Breast was the same tumor type used by Venet et al. [2], and finally (4) they represent three different tumor types. The first is Director’s Challenge (DC) dataset, which consists of 442 lung tumor samples [23] arrayed on the Affymetrix Human Genome U133A Array. This dataset was batch corrected using COMBAT [24], since it shows a clear dependence on the institution where the samples were run. The second dataset is from 286 Breast (Breast) samples (GSE2034) arrayed on the Affymetrix Human Genome U133A Array. Both of these datasets were normalized using IRON [25]. The third dataset is from The Cancer Genome Atlas (TCGA) and contains 297 primary prostate adenocarcinoma (PRAD). The level 3 Illumina HiSeq RNAseqV2 RSEM gene-level normalized mRNA expression data was downloaded from the TCGA data portal in December of 2014.

3. Results and Discussion

As illustrated in Figure 1, we propose several intrinsic characteristics of PCA-based signatures. Rather than focus on correlation of a PCA-based signature with a desired outcome, the characteristics by which a PCA-based signature may be considered valid are examined, independent of endpoint outcome.

3.1. Coherence. Individual genes in a gene signature should be correlated beyond chance, as illustrated in Figure 1(a). A coherent gene signature is an indication that a common mechanism or biological pathway is measured.

3.2. Robustness. If a gene signature describes more than one distinct biological effect, more than one PC will be significant (Figure 1(b)). PC1 may describe a combination of biological effects, but one effect should predominate. If the explained variances of PC1 and PC2 are similar to each other, this may be an indication that more than one biological effect is present. This is challenging, as the biological effect described in PC1 can easily change (PC1 and PC2 can switch ranking) if a few genes or samples are removed or if the gene signature is applied to a different dataset. Thus, it is preferred for a PCA signature to represent only a single biological effect.

3.3. Uniqueness. Datasets may be biased, meaning that many genes are not just random but actually describe a true effect such as proliferation (Venet et al. [2]). This bias can also
3.5. Measures of PCA-Based Signature Validation. Given the 4 characteristics of a PCA-based signature, we developed measures to determine the validity of these characteristics for a given signature when applied to a particular dataset (Table 1). We use the Gender-29 signature as a positive control and example; when applied to a relevant dataset (DC, or lung cancer with mixed gender), it is expected to perform well (Figure 2). Random signatures were used for comparison of PCA model parameters throughout (see Materials and Methods for details). Taken together, the validation results indicate that the Gender-29 signature is a good signature when applied to the DC dataset.

3.5.1. Coherence. For a measure of coherence, the amount of variance explained in the first principal component is used. Increased correlation among variables results in a larger
Table 1: The measures of coherence, robustness, uniqueness, and transferability applied to all datasets for all signatures used.

<table>
<thead>
<tr>
<th>Signature</th>
<th>Dataset</th>
<th>Coherence</th>
<th>Robustness</th>
<th>Uniqueness</th>
<th>Transferability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Signature PC1 (%)</td>
<td>Random mean PC1 (%)</td>
<td>Signature PC1/PC2</td>
<td>Random mean PC1/PC2</td>
</tr>
<tr>
<td>Gender-29</td>
<td>DC</td>
<td>45.3</td>
<td>10.9</td>
<td>4.57</td>
<td>1.29</td>
</tr>
<tr>
<td>Gender-29</td>
<td>Breast</td>
<td>11.8</td>
<td>12.3</td>
<td>1.32</td>
<td>1.31</td>
</tr>
<tr>
<td>TvsN-100</td>
<td>Breast</td>
<td>46.7</td>
<td>9.3</td>
<td>5.28</td>
<td>1.29</td>
</tr>
<tr>
<td>TvsN-100</td>
<td>DC</td>
<td>58.5</td>
<td>8.0</td>
<td>9.70</td>
<td>1.27</td>
</tr>
<tr>
<td>TvsN-50/50</td>
<td>DC</td>
<td>33.3</td>
<td>8.0</td>
<td>7.09</td>
<td>1.27</td>
</tr>
<tr>
<td>TvsN-25/75</td>
<td>DC</td>
<td>38.8</td>
<td>8.0</td>
<td>3.02</td>
<td>1.27</td>
</tr>
<tr>
<td>TvsN-10/90</td>
<td>DC</td>
<td>9.4</td>
<td>8.0</td>
<td>1.40</td>
<td>1.27</td>
</tr>
<tr>
<td>Mix-29/29</td>
<td>DC</td>
<td>35.2</td>
<td>8.9</td>
<td>1.55</td>
<td>1.29</td>
</tr>
<tr>
<td>Mix-29/24</td>
<td>DC</td>
<td>32.1</td>
<td>9.1</td>
<td>1.30</td>
<td>1.28</td>
</tr>
<tr>
<td>Mix-29/19</td>
<td>DC</td>
<td>28.5</td>
<td>9.3</td>
<td>1.07</td>
<td>1.28</td>
</tr>
<tr>
<td>Mix-29/14</td>
<td>DC</td>
<td>30.7</td>
<td>9.6</td>
<td>1.22</td>
<td>1.29</td>
</tr>
<tr>
<td>Gender-29</td>
<td>PRAD</td>
<td>30.8</td>
<td>20.4</td>
<td>1.62</td>
<td>1.68</td>
</tr>
<tr>
<td>TvsN-100</td>
<td>PRAD</td>
<td>51.7</td>
<td>16.9</td>
<td>5.72</td>
<td>1.69</td>
</tr>
</tbody>
</table>
explained variance in the first component, since PCA can be seen as finding the largest eigenvalue to the correlation matrix when the variables are scaled to unit variance [20]. The explained variance for the signature PCA is compared to the distribution of the explained variance of the randomized gene signature PCA models. The coherence figure (Figure 2(a)) shows that the genes in the Gender-29 signature are expressed in a coherent way, since its explained variance is 45.3%. This is larger than any of the randomized gene signature PCA models (mean: 10.9%), and none of the 10,000 random models score higher than the gender signature model in coherence.

3.5.2. Robustness. Our measurement of robustness is simple: the ratio of explained variance between PC1 and PC2. This value should be as large as possible and should also be compared to the distribution of ratios from the randomized gene signature PCA models. The results for the robustness figure (Figure 2(b)) indicate the Gender-29 signature has a PC1/PC2 explained variance ratio of 4.57. This ratio is also higher than that for the randomized gene-set PCA models. This indicates that the gender signature clearly describes a single biological effect in this dataset.

3.5.3. Uniqueness. We use PC1 from a PCA model using all the genes in the dataset as a representation of the overall direction of the dataset. The uniqueness value is derived by calculating the absolute value of the correlation between the true gene signature PCA scores versus the PCA scores of the all-gene model. This is then compared to the distribution of the absolute value correlation between the random gene signature PCA models and the PCA model using all genes. This plot indicates if there is a major variability in the dataset (most random models highly correlated with PCA scores from the all-gene model). This could be a potential problem if this major variability is also correlated to outcome, as was shown by Venet et al. [2].

The uniqueness figure (Figure 2(c)) shows that the Gender-29 signature also differentiates from the general direction of this dataset.

3.5.4. Transferability. The measurement of transferability is the correlation of the PCA loadings for the gene signature versus the reference values. If these are correlated, it implies that the gene signature describes the same biology within the dataset that it was intended to. The PCA loadings are directly related to the importance and directionality of each variable relative to the principal component [20]. This is a direct result from the fact that the scores can be seen as a weighted average of all the variables. The gender signature also describes the same biology (gender) in this dataset, since the transferability figure (Figure 2(d)) shows that the same genes have a positive loading value in both the PCA model and the reference value, and the same is true for the negative values.

In conclusion, the gender signature, when applied to the DC dataset, fulfills all the criteria for being a good signature and describes the correct biology.

3.6. The Gender Signature Fails to Translate in the Breast Dataset (Good Signature/Nonrelevant Dataset). Figure 3 shows the results for the Gender-29 signature when applied to the Breast dataset, which only contains tumor samples from females. It is clear that this signature does not work as intended for the Breast dataset. The coherence figure (Figure 3(a)) shows a lower value of 11.8% explained variance for PC1, below the random mean of 12.3%. The robustness ratio (Figure 3(b)) is also worse, at 1.32, compared to 4.57 in the DC dataset. More importantly, the uniqueness plot (Figure 3(c)) shows that its direction is similar to most of the randomized gene signature PCA models and is correlated to the general direction of the dataset. The transferability plot (Figure 3(d)) confirms that this is not a good dataset for the Gender-29 signature, since there is no correlation between the PCA loading and the reference values. All these results also show that random models can be as good as the true model.

3.7. Tumor versus Normal Signature Validation in Breast Dataset (Good Signature/Relevant Dataset). To demonstrate that valid gene signatures exist within the Breast dataset, a tumor versus normal (TvsN-100) signature, derived in lung cancer, was applied to the Breast dataset. The results clearly demonstrate that this is a good signature applied to a relevant dataset (Figure 4). The coherence is high with values of 46.7%, clearly higher than any of the 10,000 randomized gene signature PCA models. The PCA model is also robust with a PC1/PC2 ratio of 5.28, which is higher than any of the random signature PCA models. The direction of the TvsN-100 signature is also unique, with a correlation coefficient of $r = 0.036$ to the general direction. Finally, the same biology, genes that differentiate tumor samples from normal samples, is also present in this dataset, as shown by the transferability. The results are similar to the results seen for the Gender-29 signature when applied to the DC dataset (Figure 2).

3.8. Noise Injection Demonstrates the Power of PCA-Based Signatures. To investigate the effect random noise has on a PCA-based signature, we gradually increased the number of nonrelevant probesets in the TvsN-100 signature. The original TvsN-100 signature performed well in the DC dataset, as seen in Table 1 or in Supplementary Figure 1, in Supplementary Material available online at https://doi.org/10.1155/2017/2354564. The new signatures consisted of 50/50, 25/75, and 10/90 original and unrelated probesets. As can be seen in Table 1, both the TvsN-50/50 and the TvsN-25/75 signatures exhibit good statistics. This is also confirmed in the validation plots (Supplementary Fig. S2–S4). It is not until there are only 10 relevant probesets and 90 nonrelevant probesets, TvsN-10/90, that the signature starts to fall apart. This is most visible in the robustness measure, with a PC1/PC2 ratio of 1.4, and with 19% of the random PCA models having a better PC1/PC2 ratio. This is further confirmed by calculating the correlation between the original TvsN signature and the modified ones. The
correlation to the original TvsN signature is as follows: $r^2 = 0.993$, $r^2 = 0.979$, and $r^2 = 0.873$ for the TvsN-50/50, TvsN-25/75, and TvsN-10/90, respectively (Supplementary Fig. S5). It is also noteworthy to see how the explained variance falls from 58.5% for the original TvsN signature to just 9.4% for the TvsN-10/90 signature. This is expected, since more and more noise is added that is not explained by the first PCA component.

3.9. PCA Signatures Do Not Represent Mixed Signatures Robustly (Bad Signature/Relevant Dataset). As was seen in Figure 2 and Supplementary Figure 1, both the Gender-29 and the TvsN-100 signature performed well in the DC dataset. When they are merged into the mixed signature, Mix-29/29, the results are different, as can be seen in Figure 5. The coherence is still high, with an explained variance of 35.2%, higher than any of the randomized PCA models. That there are some problems becomes clear in the robustness plot (Figure 5(b)), which shows a lower PC1/PC2 ratio (1.55 versus 4.57 and 9.7 for the individual signatures) than before. This is due to the fact that the mixed signature model is not a one-component PCA model, but, rather, there are two significant principal components for this signature in this dataset. This is confirmed by investigation of the explained variance for each principal component. The explained variance for the first five PCA components are (1) 35.5%, (2) 22.7%, (3) 5.4%, (4) 3.5%, and (5) 2.7%. These numbers indicate that there are two significant principal components, since there is a large drop between components 2 and 3, but not between any of the later components. This is in accordance with the SCREE test for deciding the number of principal components [1]. That
the Mix-29/29 signature is not optimal is further exemplified in the transferability plot (Figure 5(d)), where there is no indication that the loadings from the mixed signature PCA model match the reference values.

The question then arises: what does the mix signature describe: Gender-29, TvsN-100, or a mix of both of these? Exploring the correlation between the different PCA models shows that the first PCA component is the TvsN-100 signature, $r^2 = 0.986$, and that the second PCA component is the Gender-100 signature, $r^2 = 0.989$. This also implies that, for this mixed signature, the first component is not a mix of the two signatures but is, instead, predominately one of them.

3.10. Investigation of the Stability for the Mixed Signature.
When two PCA components have similar explained variance, this can cause several problems. To demonstrate this, we made small modifications to the Mix-29/29 signature by gradually removing TvsN-100 probesets, making the gender signature more prominent. The explained variance (Supplemental Fig. S6–S8) ranges between 28.5% and 35.2% for the different signatures, with the Mix-29/29 mix model having the largest explained variance. The robustness ratio is lowest for the Mix-29/19 signature PCA model, with a PC1/PC2 ratio of 1.1. This indicates that the two signatures are almost equally important for this signature. One can also see a change in the uniqueness correlation, where the Mix-29/14 signature differs from the rest: 0.087 compared to 0.331, 0.338, and 0.255. All of these findings are further confirmed in Figure 6, where the PCA scores for the different signatures are plotted against each other. The Mix-29/24 signature is describing the same biology as the original mixed signature, since the PCA scores show a high correlation (Figure 6(a)). Figure 6(b) indicates that the Mix-29/19 signature is actually a mix of both
Figure 5: Validation plot for Mix-29/29 signature applied to DC dataset. Merging two “good” signatures and applying the resulting signature to a relevant dataset do not mean that the new signature will be valid. The probesets in the Mix-29/29 signature have a coherent expression (a), but the ratio of PC1/PC2 (b) is lower than that in the individual signatures. The uniqueness is also slightly worse (c). That the Mix-29/29 model is not working as expected is clearly seen in the transferability plot (d), where the PCA model for the DC dataset is not related to reference values.

3.11. Additional Validation on TCGA Dataset. To further confirm our finding, we also repeated the analysis for the Gender-29 and the TvsN-100 signature on a prostate adenocarcinoma (PRAD) dataset retrieved from TCGA. The results from these two signatures are shown in Table 1 and Supplementary Figures 9-10. It is clear from Table 1 and Supplementary Figure 9 that the Gender-29 signature does not work for the PRAD dataset. This is especially clear from the low PC1/PC2 explained variance ratio of 1.62 and that there is no correlation to the reference values, $r^2 = 0.025$. The results are completely opposite for the TvsN-100 signature, as can be seen in Table 1 and Supplementary Figure 10. The TvsN-100 signature shows both much higher explained variance and much higher PC1/PC2 explained variance ratio than any of the random models. Furthermore, the transferability is high, with $r^2 = 0.904$.

4. Conclusions

We have, in this manuscript, described the characteristics of PCA-based gene-expression signatures. Using the proposed
Figure 6: Correlation between the PCA scores from different mixed-gene signatures and the Gender-29 and TvsN-100 signature applied to the DC dataset. Even when five TvsN probesets are removed, the Mix-29/24 model is still correlated with the TvsN-100 signature (a). This correlation further decreases when more probesets are removed (b), until it is finally gone (c). The Mix-29/14 signature is, instead, correlated with the Gender-29 signature (d). The Gender-29 signature is less correlated with the Mix-29/19 signature (e) and shows no correlation with the Mix-29/24 signature (f).
characteristics, a signature can be validated before survival analysis or any other type of predictive modeling is done. It is important to stress the importance of validation of the signature, independent of the correlation of the signature to an outcome (e.g., survival), as was clearly shown in Venet et al. [2]. Too often, statistical significance of the signature with an outcome is used as the criteria for whether or not the signature “worked.” This is, of course, not a replacement for real biological testing that the signature is accurately predicting what it is intended to. We see this as a set of minimal requirements that any PCA-based gene signature must fulfill before moving forward with the signature in a given dataset.

Using PCA to summarize the expression of several genes has proven to be useful by others, and we also show, herein, that it is stable to random noise. However, the same properties of PCA that contribute to this usefulness can also potentially lead to misinterpretation of the signature, as shown by Venet et al. [2]. One of the most important findings presented here is that complex signatures, signatures describing multiple events, do not work well with PCA. There is always a temptation to include more genes in a signature, in order to limit the effects of outlier genes, as well as thinking that including more biologically relevant genes should result in a more stable signature. The results presented show quite the opposite, that when using PCA, a complex signature is actually less robust. The PCA will describe just one of the biological events, and which one is represented can change from dataset to dataset (Figure 6). Only if one has a perfect balance will there be a mix of the signatures (Figure 6(b)). This result is not surprising if one considers the properties of PCA. The first PCA component describes the direction in the dataset with the largest variation. The second PCA component is orthogonal to the first one and describes the second largest direction. If there are multiple biological events presented in the signature and they are not related to each other, they will thus end up in individual PCA components. The first PCA component will generally not be a combination of multiple biological events. The robustness plot addresses this by looking at the ratio of explained variance between PCI and PC2. There are, of course, many ways to estimate the number of significant principal components, many reviewed in references [26, 27], but the ratio of PCI/PC2 clearly indicates if the first PCA component is describing more explained variance than the second one.

Another important finding is that it is necessary to verify that the gene signature describes the same biology, when applied to new datasets, as it was derived for: in other words, that is it transferable. In many cases, the signature is derived from cell line experiments where something has been perturbed. The genes that show a significant change between the control and perturbed cell lines are then used as a signature. One important note is that, in a cell line system, there is much less variation compared to what is seen in, for example, tumor data. In the cell line experiment, everything is controlled and there is only one cell type. On the other hand, in tumor data, there will be much more variation from different cell types, intracellular signaling, and immune response, to name a few. A gene selected from the cell line experiment may have a very distinct expression pattern, but, in a tumor, the expression may be dependent on many more biological effects not present in the cell line experiment. This was also shown in the paper by Venet et al. [2], where many of the tested signatures were correlated to cell proliferation, even if they were derived for describing other types of biology. It has been shown, especially in breast tumors, that one of the strongest signals in tumor expression data is proliferation [2, 28, 29]. It is then easy to see that if a signature describes a weak signal, or is not distinctive, the PCA model will detect the proliferation effect in that signature as the first PCA component. This can be easily spotted by comparing the expression pattern in the cell line experiments with those seen in the tumor data, hence the need for the transferability plot. If there is no correlation seen in this plot, it indicates that when applied to a dataset, it does not work as expected. This could be due to several reasons, such as that it may be a good signature, but it was applied to a nonrelevant dataset, for example, the gender signature on the Breast dataset. Using the reference values also solves the problem with sign-flipping between PCA models.

The uniqueness plot is another indicator of if the PCA-based gene signature describes the general direction in the dataset. This is an issue if the same general direction is also predictive of outcome, as it was in the case presented by Venet et al. [2]. They showed that any random signature was as good as most of the real signatures in survival analysis. This does not mean that the signature is not working, but one cannot claim that the biology it represents is predictive, since any random signature would also be as predictive. This measure of uniqueness can be extended to include a set of validated and distinct gene-sets describing major effects seen in tumor biology, such as proliferation, epithelial-to-mesenchymal transition (EMT), or immune response. With such an analysis, it would be possible to see how the new gene signature compares to already available and validated signatures.

Lastly, the coherence plot addresses the fact that the genes within a gene signature should be expressed in a coherent way when applied to a dataset. In the derivation of a gene signature, this is commonly one of the criteria used to select the genes for the signature. Examples include selected genes that are correlated to EC50 values from the NCI-60 data or genes that are coherently expressed across several different conditions. If this is not true, when applied to another dataset, this is manifested by a low and similar explained variance compared to the randomized gene signature PCA models. A low value of the explained variance can also indicate that there are many nonrelevant genes in the signature.

We also feel that the use of a large set of randomized gene signatures enhances the results. Many of the PCA statistics used here, like explained variance, are dependent on many factors, including sample size, number of genes used, and the general behavior of the dataset. By comparing the results from the true gene signature PCA model to the results from the randomized gene signature PCA models, this problem is minimized. It also directly addressed the question if the true gene signature is better than any random signature. It is also true that a random model can be as good as the true signature,
seen in Figure 3 in this article and also in the study be Venet et al. [2].

The proposed methodologies do not remove the need for the usual best practices when it comes to using PCA to analyze data. Use score plots to find outliers, groupings, and other trends in the data that are not from biological variation, such as RNA quality and other types of batch effects. Use loading plots to see if all probesets are important for the first component, or if only some are important for later components. Furthermore, validation of survival analysis $p$ values using randomized random models is also recommended, as recently pointed out by Brulard and Chibon [30] or Venet et al. [2].

**Abbreviations**

PCA: Principal Component Analysis  
DC: Director's Challenge  
TvsN: Tumor versus normal  
PCI: First principal component  
PC2: Second principal component  
PRAD: Prostate adenocarcinoma.

**Competing Interests**

The authors declare that they have no competing interests.

**Authors’ Contributions**

Anders E. Berglund developed and implemented algorithms, performed analysis and interpretation of data, and wrote the manuscript. Eric A. Welsh and Steven A. Eschrich consulted on interpretation of data and drafting of the manuscript. Eric A. Welsh and Steven A. Eschrich contributed to drafting of the manuscript and final approval.

**Acknowledgments**

This work was supported by the National Cancer Institute [SP30-CA076292-12 to Anders E. Berglund, Eric A. Welsh, and Steven A. Eschrich] and National Cancer Institute [P50-CA119997 to Eric A. Welsh and Steven A. Eschrich].

**References**


Research Article

Module Anchored Network Inference: A Sequential Module-Based Approach to Novel Gene Network Construction from Genomic Expression Data on Human Disease Mechanism

Annamalai Muthiah,1 Susanna R. Keller,2 and Jae K. Lee1,3,4

1Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22904, USA
2Department of Medicine, Division of Endocrinology and Metabolism, University of Virginia, Charlottesville, VA 22908, USA
3Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, 12902 Magnolia Drive, Tampa, FL 33612, USA
4Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, VA 22908, USA

Correspondence should be addressed to Annamalai Muthiah; am2ta@virginia.edu

Received 1 October 2016; Accepted 15 December 2016; Published 18 January 2017

Academic Editor: Lam C. Tsoi

Copyright © 2017 Annamalai Muthiah et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Different computational approaches have been examined and compared for inferring network relationships from time-series genomic data on human disease mechanisms under the recent Dialogue on Reverse Engineering Assessment and Methods (DREAM) challenge. Many of these approaches infer all possible relationships among all candidate genes, often resulting in extremely crowded candidate network relationships with many more False Positives than True Positives. To overcome this limitation, we introduce a novel approach, Module Anchored Network Inference (MANI), that constructs networks by analyzing sequentially small adjacent building blocks (modules). Using MANI, we inferred a 7-gene adipogenesis network based on time-series gene expression data during adipocyte differentiation. MANI was also applied to infer two 10-gene networks based on time-course perturbation datasets from DREAM3 and DREAM4 challenges. MANI well inferred and distinguished serial, parallel, and time-dependent gene interactions and network cascades in these applications showing a superior performance to other in silico network inference techniques for discovering and reconstructing gene network relationships.

1. Introduction

Many established algorithms and approaches are available for inferring gene regulatory networks from large time-course molecular data [1, 2]. In silico network inference challenges under the Dialogue on Reverse Engineering Assessment and Methods (DREAM) projects—DREAM3, DREAM4, and DREAM5—have explored the strengths and weaknesses of important and widely used network inference techniques based on gene expression data. Until recently, in collaboration with the Gene Pattern team at the Broad Institute, the DREAM challenge team had selected successful network inference approaches and made them available as user friendly software algorithms and pipelines of applications that allowed users to combine multiple network inference methods on a platform so-called Gene Pattern-Dialogue on Reverse Engineering Assessment and Methods (GP-DREAM) [2, 3]. Some of the widely used network inference approaches are ANOVerence (which detects gene relationships using nonlinear correlation coefficient derived from an analysis of variance (ANOVA) [4]), correlation (which is based on pairwise correlation between genes [2]), CLR (Context Likelihood of Relatedness, which estimates gene relationships using the concept of mutual information between genes [5]), GENIE3 (which predicts expression profile of each novel gene from expression profiles of Transcription Factors using a tree based ensemble method [6]), Inferelator (network inference approach combining two key time-series data
techniques for network inference: time-lagged CLR (tCLR), an extension of CLR described above, and linear ODE model constrained by LASSO [7]), and TIGRESS (Trustful Inference of Gene Regulation using Stability Selection, a LASSO-based regression approach for inferring gene regulations [8]). Most of these network inference approaches adopt a “global” approach to network inference and construct a network using all genes simultaneously. While it is useful for initial gene network inference, such an approach often produces a “hairball-like” network that makes it hard to discern trustworthy network features among candidate network connections. Similar to the dynamic algorithm in sequence alignment, a localized approach anchoring network inference around building blocks (modules) [1] or subunits of a large network can dramatically enhance computational network reconstruction. Based on this principle, we developed our Module Anchored Network Inference (MANI) technique, which identifies gene interactions and regulatory relationships within each local module and then gradually expands the network by adding new network interactions from adjacent connected modules. This systematic and local approach to network inference constructs a less complex network and identifies dynamic relationships between network genes.

We applied MANI to time-course gene expression data of a 7-gene network during adipocyte differentiation (adipogenesis) [9]. We also tested MANI’s ability to infer two small size (n = 10) in silico networks based on perturbation time-series data from the DREAM3 and DREAM4 challenges and compared the performance of MANI against contemporary network inference methods such as ANOVerence, CLR, and TIGRESS.

2. Methods

2.1. MANI Approach. The goal of MANI is to locally infer gene regulatory relationships with sequential blocks (modules), each containing three genes (shown as a metaphorical window in Figure 1). Our three-gene module network reconstruction approach is based on (i) the observation that majority of regulatory network relationships can be captured by one of the four structures in Figure 2, each of which can be gradually reconstructed with sequential three-gene modules, and (ii) computational network inference can be efficiently performed for all possible relationships within a three-gene module. The approach consists of four steps (Figure 1): (i) identifying a set of three closest genes for the initial window(s), (ii) fitting the best possible mechanism of regulation among the genes within the initial window(s), (iii) migrating the window to next module including a new network gene which has the closest and statistically significant association with the three genes in the previous module and whose relationship has not yet been reconstructed (removing the least associated gene with the new gene among the previous three), and (iv) fitting regulatory relationships between the new gene and the genes retained from the previous module. The last two steps are repeated until all the genes potentially in the network are examined and their relationships are reconstructed if they are determined to be valid network connections based on preset criteria.

![Figure 1: Schematic of MANI steps. Step 1: selection of three genes, A, B, and C, for the initial window. Step 2: inferring regulatory relationship among genes within the window as indicated by arrows. External input is shown by U. Step 3: migrating window to accommodate a new gene (D) with the closest expression association with two genes (genes B and C) and their associated regulatory relationships from the previous window. Step 4: inferring regulatory relationships between the genes in the new window. Steps 3 and 4 are repeated until all the genes in the network are included in a window at least once.](image1)

![Figure 2: Possible gene regulatory relationships within a three-gene module. The three genes within a MANI module are labeled A, B, and C. The arrows between genes indicate the directions of regulation tested between the genes. External inputs regulating expression of genes are given as U, U₁, or U₂.](image2)

2.2. Inference of Regulatory Relationship within a Module. A local network module that contains the three most strongly correlated genes was identified by evaluating spearman rank correlations from time-series gene expression profiles. Regulatory relationships between genes within the module are inferred by selecting the optimal gene relationships from a list of possible regulatory relationships (Figure 2).

Regulatory gene relationships are mathematically modeled and fitted to gene expression data and the optimal relationship is identified using the goodness of fit measure.
the regulatory relationship of genes in a module and the number of time points in the gene expression data (n) that is used to estimate the parameters of the model [10].

\[
BIC = -2 \cdot \ln(L) + p \cdot \ln(n) \tag{4}
\]

$L$ is the maximized likelihood function of the model describing the regulatory relationship of genes. Summing up the BIC values of all genes in the module (genes A, B, and C in Figure 2), the full BIC score of any regulatory relationship in Figure 2 is

\[
BIC = \left( \frac{\text{SSE}_A}{\tau_A} + \frac{\text{SSE}_B}{\tau_B} + \frac{\text{SSE}_C}{\tau_C} \right) + p_A \cdot \ln(n_A) + p_B \cdot \ln(n_B) + p_C \cdot \ln(n_C) \tag{5}
\]

$\text{SSE}_A$, $\text{SSE}_B$, and $\text{SSE}_C$ represent the SSE of the fitted model for each gene by the regulatory relationship, $\tau_A$, $\tau_B$, and $\tau_C$ represent the standard deviation of error distribution in the fitted model of each gene. The derivation of Likelihood (L) in terms of SSE for each gene (SSE$_A$, SSE$_B$, and SSE$_C$) is shown in Supplementary Material (available online at https://doi.org/10.1155/2017/8514071). Since the number of time points in expression data is the same for each gene, $n_A = n_B = n_C = n$. Similarly, since the variance of error for each gene is also approximated to be the same, $\tau_A = \tau_B = \tau_C = \tau$. Therefore, the BIC score for a regulatory relationship is

\[
BIC = \left( \frac{\sum_{i=1}^{g} \text{SSE}_i}{\tau} \right) + p_{total} \cdot \ln(n) \tag{6}
\]

$\sum_{i=1}^{g} \text{SSE}_i$ was Sum of Square of Error (SSE) of fit for all the genes within a module (g refers to the number of genes within the module) and $p_{total} = p_A + p_B + p_C$ represents the total number of parameters in the mathematical model describing the regulatory relationship between genes.

3. Results

3.1. Implementation of MANI towards Network Inference of a 7-Gene Adipogenesis Network. The MANI approach was implemented on time-series gene expression data obtained from a network of seven genes that belong to an adipogenesis regulatory network [11]: Kruppel Like Factor 4 (KLF4), CCAAT/Enhancer Binding Protein-alpha (CEBPa), CCAAT/Enhancer Binding Protein-beta (CEBPb), CCAAT/Enhancer Binding Protein-gamma (CEBPg), GLUCose Transporter type 4 (GLUT4), Xanthine Dehydrogenase (XDH), and Peroxisome Proliferator-Activated Receptor-gamma (PPARg) (Figure 4(a)). Gene expression data had been collected during differentiation of 3T3-L1 preadipocytes into mature adipocytes for a period of 28 days [9].

Step 1 (selecting initial window(s)). The first two genes in initial windows were selected as the pair(s) of genes with maximum correlation between time-series expression data. A third gene was added by choosing a gene with maximum correlation with either of the genes forming the pair. Among the seven genes (Figure 4(a)), two pairs of genes (pair #1 = (XDH,
Figure 4: Window #1 network inference. (a) Time-series gene expression data of 7 genes within the adipogenesis network collected at 0, 6, 12, 24, 48, 72, 96, and 672 hours during adipocyte differentiation [9]. GEO accession number of gene expression data is GSE6795. Gene expression values were normalized to a maximum value of 1. Expanded view of gene expression data between 0 and 100 hours is shown in Supplementary Figure S1. (b) Time-series expression profiles of the three genes selected from the pool of 7 genes for window #1. (c) The parameters of different Regulatory Relationships (RRs) were fitted to the time-series data of genes shown in (b). Since KLF4 was determined to be the gene at the top of the hierarchy, the two remaining genes were fitted in parallel or serial regulatory relationships. Parameters describing the different regulatory relationships are as described in Figure 3. BIC scores of regulatory relationships were estimated as \( \frac{(\text{SSE}_{\text{KLF4}} + \text{SSE}_{\text{XDH}} + \text{SSE}_{\text{CEBPb}})}{\tau + p_{\text{total}}} \cdot \ln(n) \). Parallel regulatory relationship (RR #1) was chosen as optimal to describe gene relationships in window #1 because of its smallest BIC score. The parameters estimated for the selected regulatory relationship are listed in Table 1.

Table 1: Values of kinetic parameters for regulatory relationship in window #1 (RR #1 in Figure 4(c)) obtained by fitting mathematical model ((1), (2), and (3)) to gene expression data in Figure 4(b).

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Mean ± standard error (hr(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D )</td>
<td>( 0.12 ± 0.02 )</td>
</tr>
<tr>
<td>( E )</td>
<td>( 0.1 ± 0.02 )</td>
</tr>
<tr>
<td>( k_1 )</td>
<td>( 0.13 ± 0.05 )</td>
</tr>
<tr>
<td>( k_2 )</td>
<td>( 0.28 ± 0.04 )</td>
</tr>
<tr>
<td>( k_3 )</td>
<td>( 0.24 ± 0.05 )</td>
</tr>
<tr>
<td>( k_4 )</td>
<td>( 0.03 ± 0.01 )</td>
</tr>
<tr>
<td>( k_5 )</td>
<td>( 0.02 ± 0.01 )</td>
</tr>
<tr>
<td>( U )</td>
<td>( 0.06 ± 0.03 )</td>
</tr>
</tbody>
</table>

CEBPb) and pair \( #2 = (\text{CEBPa}, \text{CEBPg}) \) with the highest degrees of correlation, \( \rho = 0.88 \), were selected (correlation matrix between genes is shown in Supplementary Table S1). Following our criteria outlined above, a third gene was added to each of the two pairs resulting in window #1 = (KLF4, XDH, and CEBPb) and window #2 = (CEBPa, CEBPg, and PPARg). Expression data of genes selected for window #1 is shown in Figure 4(b).

Step 2 (fitting the best regulatory relationship for genes within the initial window(s)). The possible regulatory relationships between the three genes within a window (listed in Figure 2) were tested. Prior to this, a preliminary check was conducted to determine whether the time-course expression data of genes within a window showed differences in lags. Figure 4(b) shows there were no significant differences in lags between the three genes of window #1. In the absence of lag differences, the top gene in the hierarchy was chosen by testing all three genes (KLF, XDH, and CEBPb) within the window in that position (gene \( A \)) using (1) with a single time-invariant input
The various choices of $A$ lead to the following results: SSE of fit for the different genes were $SSE_{\text{KLF4}} = 0.0488$, $SSE_{\text{XDH}} = 0.5220$, and $SSE_{\text{CEBP}} = 0.3906$. Thus, KLF4 was the best fit at the top of the regulatory relationship in window #1. Parallel and serial regulatory relationships were then tested for the other two genes in window #1 as shown in Figure 4(c). Based on estimated BIC scores, the optimal regulatory relationship for genes in window #1 was the parallel regulatory relationship (RR#1). Solving the gene regulatory network for genes in window #2 using the same approach used for window #1, we obtained the inferred network shown in Figure 5.

Step 3 (migrating the window(s) to accommodate new genes). A new gene was introduced into the initial window using a One Gene In, One Gene Out (OISO) rule. A new gene among the remaining genes outside the window with the highest correlation with any gene inside the current window was identified while the gene least correlated with the new gene was discarded. By keeping at least one gene and its associated interactions from the previous window, we limited the number of possible regulatory relationships with the new gene(s). If introducing a new gene into the window formed an earlier window, the rule was relaxed to include the gene with the next highest degree of correlation with the genes in the window. Window #1 was thus advanced by replacing gene XDH with gene CEBP as correlation of CEBP with KLF4 ($\rho = 0.72$, Supplementary Table S1) was highest and XDH was least correlated with CEBP ($\rho = 0.03$). New window #3 thus contained KLF4, CEBPb, and CEBPg. Similarly window #2 (CEBPa, CEBPg, and PPARg) was migrated to window #4 (GLUT4, PPARg, and CEBPab) by replacing CEBP with GLUT4.

Step 4 (fitting the regulatory relationships within the new window(s)). For the new genes in the newly created windows, regulatory relationships were inferred while retaining genes and their associations from previous windows. For example, in window #3, the regulatory relation of the new gene in the window, CEBPg, was tested taking into account gene relationships to KLF4 and CEBPab from window #1. The time-course expression profiles of genes in window #3 indicated a noticeable lag for CEBPg when compared to genes KLF4 and CEBPab (Figure 4(a), Supplementary Figure S1). Thus, regulatory relationships tested potential regulation of CEBPg by KLF4 and/or CEBPab. Since CEBPg was already inferred to be regulated by gene CEBPa from window #2 (Figure 5), regulatory relationships tested in window #3 included this regulatory interaction. With testing potential gene relationships in the new windows, an additional alternate relationship was also tested, the null hypothesis scenario. The null hypothesis scenario introduces no new regulatory edges between genes in the window to prevent overfitting. The new inferred regulatory relationships for windows #3 and #4 are shown in Figure 6. MANI Steps 3 and 4 were repeated until all the 7 genes in the adipogenesis network were covered at least once by the moving windows. In total, 5 windows were created and gene relationships inferred within each window are shown in Figure 6.

The cumulative adipogenesis network inferred by MANI through the 5 windows is shown in Figure 7. In addition to the gene relationships summarized from the various windows, a gene’s likely time of activation, derived from the gene’s lag observed in the time-series expression data (Supplementary Figure S1), was included in the network. Some of the inferred gene relationships were supported by the literature. KLF4 regulates the expression of CEBPab [12] and PPARg regulates the expression of GLUT4 [13]. Fibroblasts isolated from C/EBPab−/− embryos have reduced PPARg levels and do not differentiate well when exposed to hormonal inducing agents in culture [14], implying regulation of PPARg by CEBPab. Indeed later research showed that CEBPab and PPARg regulate each other’s expression in a positive feedback loop and PPARg and CEBPab act synergistically to activate expression of fat cell specific genes such as GLUT4 [11].

### 3.2. Validation of MANI Approach

An objective validation of MANI’s performance in network inference was conducted using time-series expression data made available as part of the DREAM3 challenge (Supplementary Figure S2). This data was generated by the challenge organizers by perturbing an in silico network of 10 genes derived from *E. coli* (Figure 8(a)). The correlation matrix generated between genes using the time-series data is shown in Supplementary Table S2. The network inferred after applying MANI’s gradual and module-based local network inference approach on the DREAM3 time-series perturbation data is shown in Figure 8(b). For bigger networks of genes (number of genes in the network ($N$) ≥ 10 genes), gradual network inference by MANI leads to selection of several windows of genes. In the interest of constructing a parsimonious network, the number of windows is reduced by organizing the selected windows in decreasing order for average degrees of correlation between genes within the window and choosing only those windows from the top where a new gene in the network is selected for the first time in a window. The network is then constructed by inferring relationships of genes through the chosen windows.
**Figure 6:** Windows of 7-gene adipogenesis network. All windows covering the 7-gene network are shown. Newly inferred gene interactions inside the window are indicated by broken arrows while interactions inferred from a previous window are indicated by solid arrows. Window #5 did not have any broken arrows connecting genes because no new gene relationships were inferred; the null hypothesis was the optimal regulatory relationship connecting genes. Furthermore, windows contributing gene relationships to other windows are shown by solid arrows between windows.

**Figure 7:** Dynamic adipogenesis network constructed by MANI. The two inputs of the network were $U_1$ (time-invariant constant input) and $U_2$ (sigmoid input). The network, besides representing gene relationships collected from MANI windows in Figure 6, was also organized according to the genes’ likely times of activation in the cascade. A gene’s time of activation in the cascade was derived from a gene’s lag in its time-series expression data. Based on the ranges of times of activation in the cascade, genes were grouped in appropriate time zones in the cascade (marked by dashed vertical lines). Arranging genes in such a manner enhanced the dynamical nature of the network. Gene relationships in this MANI constructed adipogenesis network that were verified using literature are indicated with a green tick mark.

Based on this principle, the DREAM3 network was inferred by MANI using 8 windows of genes chosen from an initial list of 37 windows. The complete list of MANI selected gene windows and those windows that were chosen to infer DREAM3 network are shown in Supplementary Table S3. The values of the kinetic parameters estimated for the inferred network are given in Supplementary Table S4. The inferred network was compared to the correct DREAM3 network and the accuracy of network inference by MANI was evaluated by classifying MANI inferred edges as True Positives (TPs), False Positives (FPs), True Negatives (TNs), and False Negatives (FNs) (Figure 8(c)). The performance of MANI in network inference was compared against the performance of three other contemporary network inference approaches (ANOVerence, CLR, and TIGRESS) using the same time-series data.

Since our goal was to infer a sparse network and MANI inferred 10 edges between genes, the top 10 edges inferred by
Figure 8: Comparison of MANI inferred DREAM3 network with the correct answer. (a) The 10-gene DREAM3 network that was perturbed by DREAM3 organizers to produce the time-series data. (b) Network inference by MANI. U1, U2, and U3 were external inputs. The edges inferred between genes were either stimulating (→) or inhibitory (⊣). The parameters for the different regulatory relationships followed the same convention described in Figure 3. Genes in the network were grouped according to their likely times of activation within the cascade as estimated from the durations of lag observed in their expression data. (c) The accuracy of the DREAM3 network inferred by the MANI approach in (b) was evaluated by comparing it to the correct answer shown in (a). Comparisons were based on the presence or absence of edges between genes rather than directions of interactions. Among the MANI inferred edges in (b), there were 4 TP, 7 FN, 6 FP, and 28 True Negatives (TNs).

Table 2: Network inference performance of MANI and other methods.

<table>
<thead>
<tr>
<th>Parameters of assessment</th>
<th>ANOVerence</th>
<th>CLR</th>
<th>MANI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>27.27%</td>
<td>27.27%</td>
<td>36.36%</td>
</tr>
<tr>
<td>Specificity</td>
<td>79.41%</td>
<td>79.41%</td>
<td>82.35%</td>
</tr>
<tr>
<td>PPV</td>
<td>30%</td>
<td>30%</td>
<td>40%</td>
</tr>
</tbody>
</table>

*Note:* Sensitivity was TP/(TP + FN). Specificity was TN/(TN + FP). Positive Predictive Value (PPV) was TP/(TP + FP).

Each of the methods were used for comparison. Table 2 shows the performance of our MANI approach in network inference when compared to ANOVerence and CLR. The results for TIGRESS were similar to ANOVerence and CLR. MANI outperforms the other methods in all performance criteria. PPV of MANI with 40% is better than random network prediction as the chance of obtaining 4 correct edges and 6 wrong edges by random guesswork \((\binom{11}{4} \times \binom{7}{6}) / \binom{18}{10}\) is low (0.02). The same principles were applied to construct a size 10 network (Supplementary Table S5) from the DREAM4 challenge using two sets of perturbation data (Supplementary Figure S3). The performance of MANI (sensitivity ~27%) was comparable to TIGRESS (~33%) but worse than that of CLR (~47%).

4. Discussion

Gene expression data are generated in biological experiments at an increasing rate for the purpose of studying complex
gene regulatory mechanisms and human disease mechanisms [15–17]. Collection of time-series gene expression data has become important to deduce causal regulatory relationships between genes belonging to a network [18]. A number of network inference methods to infer gene regulatory networks from time-series gene expression data have been developed. These include solving linear ODE regression models, inferring optimal regulatory relationships between genes through a combination of procedures such as variable selection and sparse network identification, using shrinkage techniques such as LASSO and SCAD [19], solving ODE regression models to obtain possible solutions by Singular Value Decomposition followed by selection of a parsimonious network, and using various multivariate modeling techniques such as robust regression [20, 21], Dynamic Bayesian Network modeling [22–24], and time-delayed ARCANE algorithm that infers gene relationships based on mutual information between genes [25]. The common feature of the currently used methods is that they generate a global network of gene relationships in an unsupervised manner. Therefore, the constructed network is often crowded and does not provide a clear delineation of the network’s hierarchical or dynamical features (e.g., Figure 8(a)). With increasing numbers of genes in the network, the complexity of interconnections between genes increases exponentially, making in silico reconstruction of the network rapidly intractable.

In contrast, MANI adopts a systematic and gradual approach to network inference by constructing networks within local modules. This local approach to network inference adopted by MANI allowed the final constructed adipogenesis network (Figure 7) to be sparse and well organized, highlighting structural aspects of the network such as the hierarchy in gene relationships and also providing clarity to the network’s pathways of activation. MANI’s inference of the adipogenesis network in Figure 7 shows that the network follows a serial-parallel pathway for cascade activation. MANI was successful in inferring a hierarchy of regulation between genes when a difference in lag was detected in gene expression profiles. Regulatory relationships between genes G3, G4, G5, G6, and G9 in the DREAM3 network (Figure 8(c)) and genes CEBPa, PPARg, and GLUT4 in the adipogenesis network (Figure 7) were successfully inferred. Adding times of activation of genes to the MANI constructed network enhanced the overall quality of the inferred network by making it more dynamic and interpretable. For example, in the case of the adipogenesis network in Figure 7, arranging genes according to their times of activation showed how the genes in the network switched on at various time intervals. Therefore, MANI inference provided a structural organization of genes in the network. The accuracy of the local approach towards constructing networks proposed by MANI was modestly better than that of other well-known global network inference methods in the DREAM3 challenge (Table 2). The reason for MANI’s lower performance in the DREAM4 challenge was due to the presence of feedback loops in the network while MANI has been primarily developed to infer networks without such feedback structures. We believe MANI’s approach still has high value in identifying novel biological networks without such loop connections. MANI distinguishes itself from other global network inference approaches in that it can locally yet dynamically reconstruct networks across moving modules and windows and can easily be extended to reconstruct much larger networks. While constructing larger networks using MANI, the inferred network is a local optimum (instead of global) since MANI infers the network using locally constructed network modules, and then additional edges in the network are gradually expanded from neighboring modules. In this regard, MANI is also one of the heuristic algorithms, following a search path based on high probability regulatory expression association of network genes.

We note that the current MANI approach also has several limitations. Inference of hierarchy in the network relies on differences in lags between the expressions of different genes. Therefore, lack of differences in lags between genes hinders MANI’s ability to infer regulatory relationships between genes. Relationships between genes G1, G2, G5, and G8 in the DREAM3 network (Figure 8(c)) were incorrectly inferred due to the lack of differences in lags between them in the time-series expression data that was used to construct the network. MANI also currently relies on a constant or sigmoid perturbation rate (U) in our ODE model for network inference, which can be relaxed in a future study. Time-series data obtained by a single perturbation of the network may also activate multiple genes within the network and, therefore, in order to maximize network inference performance by MANI, multiple time-series data generated by multiple perturbations of the same network can be used for improved network inference to distinguish such multiple interactions. Furthermore, the current ODE model of MANI is best suited for time-series gene expression data to infer gene regulatory networks. MANI’s scope can be expanded by developing a local approach to network inference using static gene expression data in future applications.

Competing Interests
The authors declare no conflict of interests.

Acknowledgments
The work described in here was supported by a grant from the National Institutes of Health [R01 HL081690]. Mr. Joe Simard at the Software Site Licensing Department at the University of Virginia Information Technology Services helped to set up the SIMBiology toolbox for use within the authors’ lab MATLAB platform. Assistance in using the toolbox to optimize parameters of ODE models was provided by MathWorks staff (Mr. Scott Benway and Ms. Josephine Dula).

References


Review Article
A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data

Jennifer D. Hintzsche,1 William A. Robinson,1,2 and Aik Choon Tan1,2,3

1Division of Medical Oncology, Department of Medicine, School of Medicine, Aurora, CO 80045, USA
2University of Colorado Cancer Center, Aurora, CO 80045, USA
3Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado, Anschutz Medical Campus, Aurora, CO 80045, USA

Correspondence should be addressed to Aik Choon Tan; aikchoon.tan@ucdenver.edu

Received 25 May 2016; Accepted 26 October 2016

Copyright © 2016 Jennifer D. Hintzsche et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Whole Exome Sequencing (WES) is the application of the next-generation technology to determine the variations in the exome and is becoming a standard approach in studying genetic variants in diseases. Understanding the exomes of individuals at single base resolution allows the identification of actionable mutations for disease treatment and management. WES technologies have shifted the bottleneck in experimental data production to computationally intensive informatics-based data analysis. Novel computational tools and methods have been developed to analyze and interpret WES data. Here, we review some of the current tools that are being used to analyze WES data. These tools range from the alignment of raw sequencing reads all the way to linking variants to actionable therapeutics. Strengths and weaknesses of each tool are discussed for the purpose of helping researchers make more informative decisions on selecting the best tools to analyze their WES data.

1. Introduction

Recent advances in next-generation sequencing technologies provide revolutionary opportunities to characterize the genomic landscapes of individuals at single base resolution for identifying actionable mutations for disease treatment and management [1, 2]. Whole Exome Sequencing (WES) is the application of the next-generation technology to determine the variations in the exome, that is, all coding regions of known genes in a genome. For example, more than 85% of disease-causing mutations in Mendelian diseases are found in the exome, and WES provides an unbiased approach to detect these variants in the era of personalized and precision medicine. Next-generation sequencing technologies have shifted the bottleneck in experimental data production to computationally intensive informatics-based data analysis. For example, the Exome Aggregation Consortium (ExAC) has assembled and reanalyzed WES data of 60,706 unrelated individuals from various disease-specific and population genetic studies [3]. To gain insights in WES, novel computational algorithms and bioinformatics methods represent a critical component in modern biomedical research to analyze and interpret these massive datasets.

Genomic studies that employ WES have increased over the years, and new bioinformatics methods and computational tools have developed to assist the analysis and interpretation of this data (Figure 1). The majority of WES computational tools are centered on the generation of a Variant Calling Format (VCF) file from raw sequencing data. Once the VCF files have been generated, further downstream analyses can be performed by other computational methods. Therefore, in this review we have classified bioinformatics methods and computational tools into Pre-VCF and Post-VCF categories. Pre-VCF workflows include tools for aligning the raw sequencing reads to a reference genome, variant detection, and annotation. Post-VCF workflows include tools for somatic mutation detection, pathway analysis, copy number alterations, INDEL identification, and driver prediction. Depending on the nature of the hypothesis, beyond VCF
bioinformatics and computational tools as a reference for WES studies (Table 1).

2. Computational Tools in Pre-VCF Analyses

Alignment, removal of duplicates, variant calling, annotation, filtration, and prediction are all parts of the steps leading up to the generation of a filtered and annotated VCF file. Here we review each one of these steps, as shown in Figure 2, and compare and contrast some of the tools that can be used to perform the Pre-VCF analysis steps.

2.1. Alignment Tools. The first step in any analysis of next-generation sequencing is to align the sequencing reads to a reference genome. The two most common reference genomes for humans currently are hg18 and hg19. Several aligning algorithms have been developed including but not limited to BWA [8], Bowtie 1 [9] and 2 [10], GEM [11], ELAND (Illumina, Inc.), GSNAP [12], MAQ [13], mrFAST [14], Novoalign (http://www.novocraft.com/), SOAP1 [15] and 2 [16], SSAHA [17], Stampy [18], and YOABS [19]. Each method has its own unique features and many papers have reviewed the differences between them [20–22], and we will not review these tools in depth here. The three most commonly used of these algorithms are BWA, Bowtie (1 and 2), and SOAP (1 and 2).

2.2. Auxiliary Tools. Some auxiliary tools have been developed to filter aligned reads to ensure higher quality data for downstream analyses. PCR amplification can introduce duplicate reads of paired-end reads in sequencing data. These duplicate reads can influence the depth of the mapped reads and downstream analyses. For example, if a variant is detected in duplicate reads, the proportion of reads containing a variant could pass the threshold needed for variant calling, thus calling a falsely positive variant. Therefore, removing duplicate reads is a crucial step in accurately representing the sequencing depth during downstream analyses. Several tools have been developed to detect PCR duplicates including Picard (http://picard.sourceforge.net/), FastUniq [23], and SAMtools [7]. SAMtools rmdup finds reads that start and end at the same position, find the read with the highest quality score, and mark the rest of the duplicates for removal. Picard finds identical 5’ positions for both reads in a mate pair and marks them as duplicates. In contrast, FastUniq takes a de novo approach to quickly identify PCR duplicates. FastUniq imports all reads, sorts them according to their location, and then marks duplicates. This allows FastUniq to not to require complete genome sequences as prerequisites. Due to the different algorithms each of these tools use, these tools can remove PCR duplicates individually or in combination.

2.3. Methods for Single Nucleotide Variants (SNVs) Calling. After sequences have been aligned to the reference genome, the next step is to perform variant detection in the WES data. There are four general categories of variant calling strategies: germline variants, somatic variants, copy number variations, and structural variants. Multiple tools that perform one or more of these variant calling techniques were recently compared to each other [24]. Some common SNV calling
programs are GATK [4–6], SAMtools [7], and VCMM [25]. The actual SNV calling mechanisms of GATK and SAMtools are very similar. However, the context before and after SNV calling represents the differences between these tools. GATK assumes each sequencing error is independent while SAMtools believes a secondary error carries more weight. After SNV calling GATK learns from data while SAMtools relies on options of the user. Variant Caller with Multinomial probabilistic Model (VCMM) is another tool developed to detect SNVs and INDELs from WES and Whole Genome Sequencing (WGS) studies using a multinomial probabilistic model with quality score and a strand bias filter [25]. VCMM suppressed the false-positive and false-negative variant calls when compared to GATK and SAMtools. However, the number of variant calls was similar to previous studies. The comparison done by the authors of VCMM demonstrated that while all three methods call a large number of common SNVs, each tool also identifies SNVs not found by the other methods [25]. The ability of each method to call SNVs not found by the others should be taken into account when choosing a SNV variant calling tool(s).

2.4. Methods for Structural Variants (SVs) Identification. Structural Variants (SVs) such as insertions and deletions (INDELs) in high-throughput sequencing data are more challenging to identify than single nucleotide variants because they could involve an undefined number of nucleotides. The majority of WES studies follow SAMtools [7] or GATK [4–6] workflows which will identify INDELs in the data. However, other software has been developed to increase the sensitivity of INDEL discovery while simultaneously decreasing the false discovery rate.

Platypus [26] was developed to find SNVs, INDELs, and complex polymorphisms using local de novo assembly. When compared to SAMtools and GATK, Platypus had the lowest Fosmid false discovery rate for both SNVs and INDELs in whole genome sequencing of 15 samples. It also had the shortest runtime of these tools. However, GATK and SAMtools had lower Fosmid false discovery rates than Platypus when finding SNVs and INDELs in WES data [26]. Therefore, Platypus seems to be appropriate for whole genome sequencing but caution should be used when utilizing this tool with WES data.

FreeBayes uses a unique approach to INDEL detection compared to other tools. The method utilizes haplotype-based variant detection under a Bayesian statistics framework [27]. This method has been used in several studies in combination with other approaches for the identifying of unique INDELs [28, 29].

Pindel was one of the first programs developed to address the issue of unidentified large INDELs due to the short length of WGS reads [30]. In brief, after alignment of the reads to the reference genome, Pindel identifies reads where one end was mapped and the other was not [30]. Then, Pindel searches the reference genome for the unmapped portion of this read over a user defined area of the genome [30]. This split-read algorithm successfully identified large INDELs. Other computational tools developed after Pindel still utilize this algorithm as the foundation in their methods for detecting INDELs.

Splitread [31] was developed to specifically identify structural variants and INDELs in WES data from 1bp to 1Mbp building on the split-read approach of Pindel [30]. The algorithms used by SAMtools and GATK limit the size of

Figure 2: Whole Exome Sequencing data analysis steps. Novel computational methods and tools have been developed to analyze the full spectrum of WES data, translating raw fastq files to biological insights and precision medicine.
<table>
<thead>
<tr>
<th>Computational tools</th>
<th>Description</th>
<th>Website</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alignment tools</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burrows-Wheeler Aligner (BWA)</td>
<td>Perform short reads alignment using BWT approach against a references genome allowing for gaps/mismatches. Performs short read alignment using the Burrows-Wheeler index in order to be memory efficient, while still maintaining an alignment speed of over 25 million 35 bp reads per hour.</td>
<td><a href="http://bio-bwa.sourceforge.net/">http://bio-bwa.sourceforge.net/</a></td>
<td>[8]</td>
</tr>
<tr>
<td>Bowtie (1 &amp; 2)</td>
<td>Short read aligner that achieves speed by splitting reads into equal lengths and applying seed templates to guarantee hits with only 2 mismatches.</td>
<td><a href="http://bowtie-bio.sourceforge.net/index.shtml">http://bowtie-bio.sourceforge.net/index.shtml</a></td>
<td>[9, 10]</td>
</tr>
<tr>
<td>GSNAP</td>
<td>Short read aligner using a hashing algorithm to find exact or close to exact matching in DNA and protein databases, analogous to doing a BLAST search for each read.</td>
<td><a href="http://research-pub.gene.com/gmap/">http://research-pub.gene.com/gmap/</a></td>
<td>[12]</td>
</tr>
<tr>
<td>MAQ</td>
<td>Alignment done on paired-end or single-end sequences, also capable of doing methylation studies. Allows for a mismatch up to 50% of a read length and has built-in adapter and base quality trimming. SOAP2 improved speed by an order of magnitude over SOAP1 and can align a wide range of read lengths at the speed of 2 minutes for one million single-end reads using a two-way BWT algorithm. Uses a hashing algorithm to find exact or close to exact matching in DNA and protein databases, analogous to doing a BLAST search for each read.</td>
<td><a href="http://maq.sourceforge.net/">http://maq.sourceforge.net/</a></td>
<td>[13]</td>
</tr>
<tr>
<td>mrFAST</td>
<td>Alignment done using a hashing algorithm and statistical model, to align Illumina reads for genome, RNA, and Chip sequencing allowing for a large number or variations including insertions and deletions.</td>
<td><a href="http://mrfast.sourceforge.net/">http://mrfast.sourceforge.net/</a></td>
<td>[14]</td>
</tr>
<tr>
<td>SOAP (1 &amp; 2)</td>
<td></td>
<td><a href="http://soap.genomics.org.cn/">http://soap.genomics.org.cn/</a></td>
<td>[15, 16]</td>
</tr>
<tr>
<td>Stampy</td>
<td></td>
<td><a href="http://www.well.ox.ac.uk/project-stampy">http://www.well.ox.ac.uk/project-stampy</a></td>
<td>[18]</td>
</tr>
<tr>
<td>Computational tools</td>
<td>Description</td>
<td>Website</td>
<td>References</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------</td>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>YOABS</td>
<td>Uses a $O(n)$ algorithm that uses both hash and tri-based methods that are effective in aligning sequences over 200 bp with 3 times less memory and ten times faster than SSAHA.</td>
<td>Available by request for noncommercial use</td>
<td>[19]</td>
</tr>
<tr>
<td>HTSeq</td>
<td>Python based package with many functions to facilitate several aspects of sequencing studies.</td>
<td><a href="http://www-huber.embl.de/HTSeq/doc/overview.html">http://www-huber.embl.de/HTSeq/doc/overview.html</a></td>
<td></td>
</tr>
<tr>
<td><strong>Auxiliary tools</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FastUniq</td>
<td>Imports, sorts, and identifies PCR duplicates of short sequences from sequencing data.</td>
<td><a href="https://sourceforge.net/projects/fastuniq/">https://sourceforge.net/projects/fastuniq/</a></td>
<td>[23]</td>
</tr>
<tr>
<td>Picard</td>
<td>high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.</td>
<td><a href="http://picard.sourceforge.net/">http://picard.sourceforge.net/</a></td>
<td></td>
</tr>
<tr>
<td><strong>SNV and SV calling</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATK</td>
<td>Variant calling of SNPs and small INDELs; can also be used on nonhuman and nondiploid organisms.</td>
<td><a href="https://www.broadinstitute.org/gatk/">https://www.broadinstitute.org/gatk/</a></td>
<td>[4–6]</td>
</tr>
<tr>
<td>FreeBayes</td>
<td>Detection of SNPs, MNPs, INDELs, and structural variants (SVs) from sequencing alignments using Bayesian statistical methods.</td>
<td><a href="https://github.com/ekg/freebayes">https://github.com/ekg/freebayes</a></td>
<td>[27]</td>
</tr>
<tr>
<td>indelMINER</td>
<td>Splitread algorithm to identify breakpoint in INDELs from paired-end sequencing data.</td>
<td><a href="https://github.com/aakrosh/indelMINER">https://github.com/aakrosh/indelMINER</a></td>
<td>[32]</td>
</tr>
<tr>
<td>Pindel</td>
<td>Detection of INDELs using a pattern growth approach with anchor points to provide nucleotide-level resolution. Detection of SNPs, MNPs, INDELs, replacements, and structural variants (SVs) from sequencing alignments using local realignment and local assembly to achieve high specificity and sensitivity.</td>
<td><a href="http://gmt.genome.wustl.edu/packages/pindel/">http://gmt.genome.wustl.edu/packages/pindel/</a></td>
<td>[30]</td>
</tr>
<tr>
<td>Platypus</td>
<td>Detection of INDELs less than 50 bp long from WES or WGS data, using a split-read algorithm.</td>
<td><a href="http://www.well.ox.ac.uk/platypus">http://www.well.ox.ac.uk/platypus</a></td>
<td>[26]</td>
</tr>
<tr>
<td>Splitread</td>
<td>Detection of INDELs is done using a split-read and soft-clipping approach that is especially sensitive in datasets with low coverage.</td>
<td><a href="http://splitread.sourceforge.net/">http://splitread.sourceforge.net/</a></td>
<td>[31]</td>
</tr>
<tr>
<td>sprites</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>VCF annotation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANNOVAR</td>
<td>Provides up-to-date annotation of VCF files by gene, region, and filters from several other databases.</td>
<td><a href="http://annovar.openbioinformatics.org/">http://annovar.openbioinformatics.org/</a></td>
<td>[34]</td>
</tr>
<tr>
<td>SnpEff</td>
<td>Uses 38,000 genomes to predict and annotate the effects of variants on genes.</td>
<td><a href="http://snpeff.sourceforge.net/">http://snpeff.sourceforge.net/</a></td>
<td>[36]</td>
</tr>
<tr>
<td>SnpSift</td>
<td>Tools to manipulate VCF files including filtering, annotation, case controls, transition, and transversion rates and more.</td>
<td><a href="http://snpeff.sourceforge.net/SnpSift.html">http://snpeff.sourceforge.net/SnpSift.html</a></td>
<td>[37]</td>
</tr>
<tr>
<td>VAT</td>
<td>Annotation of variants by functionality in a cloud computing environment.</td>
<td><a href="http://vat.gersteinlab.org/">http://vat.gersteinlab.org/</a></td>
<td>[38]</td>
</tr>
<tr>
<td>Computational tools</td>
<td>Description</td>
<td>Website</td>
<td>References</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------</td>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td><strong>Database filtration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000 Genomes Project</td>
<td>Genotype information from a population of 1000 healthy individuals.</td>
<td><a href="http://www.1000genomes.org/">http://www.1000genomes.org/</a></td>
<td>41</td>
</tr>
<tr>
<td>LOVD</td>
<td>Open source database of freely available gene-centered collection of DNA variants and storage of patient and NGS data.</td>
<td><a href="http://www.lovd.nl/3.0/home">http://www.lovd.nl/3.0/home</a></td>
<td>40</td>
</tr>
<tr>
<td>COSMIC</td>
<td>Database containing somatic mutations from human cancers separated into expert curated data and genome-wide screen published in scientific literature.</td>
<td><a href="http://cancer.sanger.ac.uk/cosmic">http://cancer.sanger.ac.uk/cosmic</a></td>
<td>42</td>
</tr>
<tr>
<td>NHLBI GO Exome Sequencing Project (ESP)</td>
<td>Database of genes and mechanisms that contribute to blood, lung, and heart disorders through NGS data in various populations.</td>
<td><a href="http://evs.gs.washington.edu/EVS/">http://evs.gs.washington.edu/EVS/</a></td>
<td></td>
</tr>
<tr>
<td>SeattleSeq Annotation</td>
<td>Part of the NHBLI sequencing project; this database contains novel and known SNVs and INDELs including accession number, function of the variant, and HapMap frequencies, clinical association, and PolyPhen predictions.</td>
<td><a href="http://snp.gs.washington.edu/SeattleSeqAnnotation137/">http://snp.gs.washington.edu/SeattleSeqAnnotation137/</a></td>
<td></td>
</tr>
<tr>
<td><strong>Functional predictors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CADD</td>
<td>Machine learning algorithm to score all possible 8.6 million substitutions in the human reference genome from 1 to 99 based on known and simulated functional variants.</td>
<td><a href="http://cadd.gs.washington.edu/info">http://cadd.gs.washington.edu/info</a></td>
<td>49</td>
</tr>
<tr>
<td>LRT</td>
<td>Uses the Likelihood Ratio statistical test to compare a variant to known variants and determine if they are predicted to be benign, deleterious, or unknown.</td>
<td><a href="http://genome.cshlp.org/content/19/9/i553Jong">http://genome.cshlp.org/content/19/9/i553Jong</a></td>
<td>45</td>
</tr>
<tr>
<td>PolyPhen-2</td>
<td>Predicts potential impact of a nonsynonymous variant using comparative and physical characteristics.</td>
<td><a href="http://genetics.bwh.harvard.edu/pph2/">http://genetics.bwh.harvard.edu/pph2/</a></td>
<td>44</td>
</tr>
<tr>
<td>SIFT</td>
<td>By using PSI-BLAST, a prediction can be made on the effect of a nonsynonymous mutation within a protein.</td>
<td><a href="http://sift.jcvi.org/">http://sift.jcvi.org/</a></td>
<td>43</td>
</tr>
<tr>
<td>VEST</td>
<td>Machine learning approach to determine the probability that a missense mutation will impair the functionality of a protein.</td>
<td><a href="http://karchinlab.org/apps/appVest.html">http://karchinlab.org/apps/appVest.html</a></td>
<td>48</td>
</tr>
<tr>
<td>MetaSVM &amp; MetaLR</td>
<td>Integration of a Support Vector Machine and Logistic Regression to integrate nine deleterious prediction scores of missense mutations.</td>
<td><a href="https://sites.google.com/site/jpropgen/dbNSFP">https://sites.google.com/site/jpropgen/dbNSFP</a></td>
<td>47</td>
</tr>
<tr>
<td>Computational tools</td>
<td>Description</td>
<td>Website</td>
<td>References</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------</td>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td><strong>Significant somatic mutations</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SomaticSniper</strong></td>
<td>Using two bam files as input, this tool uses the genotype likelihood model of MAZ to calculate the probability that the tumor and normal samples are different, thus identifying somatic variants.</td>
<td>[50]</td>
<td></td>
</tr>
<tr>
<td><strong>MuTect</strong></td>
<td>Using statistical analysis to predict the likelihood of a somatic mutation using two Bayesian approaches.</td>
<td>[35]</td>
<td></td>
</tr>
<tr>
<td><strong>VarSim</strong></td>
<td>By leveraging on previously reported mutations, a random mutation simulation is performed to predict somatic mutations.</td>
<td>[51]</td>
<td></td>
</tr>
<tr>
<td><strong>SomVarIUS</strong></td>
<td>Identification of somatic variants from unpaired tissue samples with a sequencing depth of 150x and 67% precision, implemented in Python.</td>
<td>[52]</td>
<td></td>
</tr>
<tr>
<td><strong>Copy number alteration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Control-FREEC</strong></td>
<td>Detects copy number changes and loss of heterozygosity (LOH) from paired SAM/BAM files by computing and normalizing copy number and beta allele frequency.</td>
<td>[59]</td>
<td></td>
</tr>
<tr>
<td><strong>CNV-seq</strong></td>
<td>Mapped read count is calculated over a sliding window in Perl and R to determine copy number from HTS studies.</td>
<td>[53]</td>
<td></td>
</tr>
<tr>
<td><strong>SegSeq</strong></td>
<td>Using 14 million aligned sequence reads from cancer cell lines, equal copy number alterations are calculated from sequencing data. Determines copy number changes in matched or unmatched samples using read ratios and then postprocessed with a circular binary segmentation algorithm.</td>
<td>[54]</td>
<td></td>
</tr>
<tr>
<td><strong>VarScan2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ExomeAI</strong></td>
<td>Detects copy number variants from unpaired tissue samples using a statistical approach that is capable of handling low-quality datasets.</td>
<td>[64]</td>
<td></td>
</tr>
<tr>
<td><strong>CNV-seqer</strong></td>
<td>Exon coverage between matched sequences was calculated using log2 ratios followed by the circular binary segmentation algorithm.</td>
<td>[60]</td>
<td></td>
</tr>
<tr>
<td><strong>EXCAVATOR</strong></td>
<td>Detects copy number variants from WES data in 3 steps using a Hidden Markov Model algorithm.</td>
<td>[57]</td>
<td></td>
</tr>
<tr>
<td><strong>ExomeCNV</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ADTEx</strong></td>
<td>Detection of aberrations in tumor exomes by detecting B-allele frequencies and implemented in R.</td>
<td>[55]</td>
<td></td>
</tr>
<tr>
<td><strong>CONTRA</strong></td>
<td>Uses normalized depth of coverage to detect copy number changes from targeted resequencing data including WES.</td>
<td>[56]</td>
<td></td>
</tr>
<tr>
<td>Computational tools</td>
<td>Description</td>
<td>Website</td>
<td>References</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------</td>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td><strong>Driver prediction tools</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHASM</td>
<td>Machine learning method that predicts the functional significance of somatic mutations. De novo drivers are discovered from cancer only mutational data including genes, nucleotides, or domains that have high exclusivity and coverage.</td>
<td><a href="http://karchinlab.org/apps/appChasm.html">http://karchinlab.org/apps/appChasm.html</a></td>
<td>[65]</td>
</tr>
<tr>
<td>Dendrix</td>
<td>Gene-specific and patient-specific mutation frequencies are incorporated to find mutations in genes that are mutated more often than would be expected by chance.</td>
<td><a href="http://compbio.cs.brown.edu/projects/dendrix/">http://compbio.cs.brown.edu/projects/dendrix/</a></td>
<td>[66]</td>
</tr>
<tr>
<td>MutSigCV</td>
<td>Gene-specific and patient-specific mutation frequencies are incorporated to find mutations in genes that are mutated more often than would be expected by chance.</td>
<td><a href="http://www.broadinstitute.org/cancer/software/genepattern/modules/docs/MutSigCV">http://www.broadinstitute.org/cancer/software/genepattern/modules/docs/MutSigCV</a></td>
<td>[67]</td>
</tr>
<tr>
<td><strong>Pathway analysis tools and resources</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEGG</td>
<td>Database using maps of known biological processes that allows searching for genes and color coding of results. Allows for users to input a large set of genes and discover the functional annotation of the gene list including pathways, gene ontology terms, and more.</td>
<td><a href="http://www.genome.jp/kegg/">http://www.genome.jp/kegg/</a></td>
<td>[68]</td>
</tr>
<tr>
<td>DAVID</td>
<td>Network visualization of protein-protein interactions of over 2,031 organisms.</td>
<td><a href="http://string-db.org/">http://string-db.org/</a></td>
<td>[70]</td>
</tr>
<tr>
<td>STRING</td>
<td>Uses biomedical knowledge to allow users to search for relationships between biomedical entities.</td>
<td><a href="http://infos.korea.ac.kr/berex/">http://infos.korea.ac.kr/berex/</a></td>
<td>[71]</td>
</tr>
<tr>
<td>BEReX</td>
<td>Uses a list of genes to determine physical connectivity among proteins according to protein-protein interactions.</td>
<td><a href="http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001273">http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001273</a></td>
<td>[72]</td>
</tr>
<tr>
<td>DAPPLE</td>
<td>Uses linkage disequilibrium to determine pathways and cell types that are likely to be affected based on SNP data.</td>
<td><a href="http://www.broadinstitute.org/mpg/snpsea/">http://www.broadinstitute.org/mpg/snpsea/</a></td>
<td>[73]</td>
</tr>
<tr>
<td>SNPsea</td>
<td>Database using maps of known biological processes that allows searching for genes and color coding of results. Allows for users to input a large set of genes and discover the functional annotation of the gene list including pathways, gene ontology terms, and more.</td>
<td><a href="http://www.cbioportal.org/">http://www.cbioportal.org/</a></td>
<td>[78]</td>
</tr>
<tr>
<td>My Cancer Genome</td>
<td>Database for cancer research that provides linkage of mutational status to therapies and available clinical trials.</td>
<td><a href="https://www.mycancergenome.org/">https://www.mycancergenome.org/</a></td>
<td><a href="http://www.mycancergenome.org/">http://www.mycancergenome.org/</a> [74]</td>
</tr>
<tr>
<td>ClinVar</td>
<td>Database of relationship between phenotypes and human variations, showing the relationship between health status and human variations and known implications. Database of drug signatures that includes 19,531 genes and 17,389 compounds that can in part help identify compounds for drug repurposing studies in translational research.</td>
<td><a href="https://www.ncbi.nlm.nih.gov/clinvar/">https://www.ncbi.nlm.nih.gov/clinvar/</a></td>
<td>[74]</td>
</tr>
<tr>
<td>DSigDB</td>
<td>Knowledge base allowing visualization of a variety of drug-gene knowledge.</td>
<td><a href="http://tanlab.ucdenver.edu/DSigDB">http://tanlab.ucdenver.edu/DSigDB</a></td>
<td>[77]</td>
</tr>
<tr>
<td>PharmGKB</td>
<td>Contains detailed drug information with comprehensive drug target information for 8,206 drugs.</td>
<td><a href="http://www.pharmgkb.org/">http://www.pharmgkb.org/</a></td>
<td>[75]</td>
</tr>
<tr>
<td>DrugBank</td>
<td></td>
<td><a href="http://www.drugbank.ca/">http://www.drugbank.ca/</a></td>
<td>[76]</td>
</tr>
<tr>
<td>Computational tools</td>
<td>Description</td>
<td>Website</td>
<td>References</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td><strong>WES data analysis pipelines</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fast2VCF</td>
<td>Whole Exome Sequencing pipeline that starts with raw sequencing (fastq) files and ends with a VCF file that has good capability for novel and expert users. WES or WGS pipeline that combines the information from over ten alignment and analysis tools to arrive at a VCF file that can be used in both Mendelian and cancer studies.</td>
<td><a href="http://fastq2vcf.sourceforge.net/">http://fastq2vcf.sourceforge.net/</a></td>
<td>[80]</td>
</tr>
<tr>
<td>SeqMule</td>
<td>WES data analysis pipeline that starts with raw sequencing reads and analyzes SNVs and CNAs and links this data to a list of prioritized drugs from clinical trials and D8igDB.</td>
<td><a href="http://seqmule.openbioinformatics.org/en/latest/">http://seqmule.openbioinformatics.org/en/latest/</a></td>
<td>[79]</td>
</tr>
<tr>
<td>IMPACT</td>
<td>Automated sequencing pipeline that performs in part alignment, variant calling, and quality control that can be run on Amazon Web Services EC2 as well as local machines and clusters.</td>
<td><a href="http://tanlab.ucdenver.edu/IMPACT/">http://tanlab.ucdenver.edu/IMPACT/</a></td>
<td>[81]</td>
</tr>
<tr>
<td>Genomes on the Cloud</td>
<td></td>
<td><a href="http://genome.sph.umich.edu/wiki/GotCloud">http://genome.sph.umich.edu/wiki/GotCloud</a></td>
<td></td>
</tr>
<tr>
<td>(GotCloud)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
structural variants, with variants greater than 15 bp rarely being identified [31]. Splitread anchors one end of a read and clusters the unanchored ends to identify size, content, and location of structural variants [31]. When compared to GATK, Splitread called 70% of the same INDELS but identified 19 more unique INDELS, 13 of which were verified by sanger sequencing [31]. The unique ability of Splitread to identify large structural variants and INDELS merits it being used in conjunction with other INDEL detecting software in WES analysis.

Recently developed indelMINER is a compilation of tools that takes the strengths of split-read and de novo assembly to determine INDELS from paired-end reads of WGS data [32]. Comparisons were done between SAMtools, Pindel, and indelMINER on a simulated dataset with 7,500 INDELS [32]. SAMtools found the least INDELS with 6,491, followed by Pindel with 7,239 and indelMINER with 7,365 INDELS identified. However, indelMINER’s false-positive percentage (3.57%) was higher than SAMtools (2.65%) but lower than Pindel (4.53%). Conversely, indelMINER did have the lowest number of false-negatives with 398 compared to 589 and 1,181 for Pindel and SAMtools, respectively. Each of these tools has its own strengths and weaknesses as demonstrated by the authors of indelMINER [32]. Therefore, it can be predicted that future tools developed for SV detection will take an approach similar to indelMINER in trying to incorporate the best methods that have been developed thus far.

Most of the recent SV detection tools rely on realigning split-reads for detecting deletions. Instead of a more universal approach like indelMINER, Sprites [33] aims to solve the problem of deletions with microhomologies and deletions with microinsertions. Sprites algorithm realigns soft-clipping reads to find the longest prefix or suffix that has a match in the target sequence. In terms of the F-score, Sprites performed better than Pindel using real and simulated data [33].

All of these tools use different algorithms to address the problem of structural variants, which are common in human genomes. Each of these tools has strengths and weaknesses in detecting SVs. Therefore, it is suggested to use several of these tools in combination to detect SVs in WES.

2.5. VCF Annotation Methods. Once the variants are detected and called, the next step is to annotate these variants. The two most popular VCF annotation tools are ANNOVAR [34] and MuTect [35] which is part of the GATK pipeline. ANNOVAR was developed in 2010 with the aim to rapidly annotate millions of variants with ease and remains one of the popular variant annotation methods to date [34]. ANNOVAR can use gene, region, or filter-based annotation to access over 20 public databases for variant annotation. MuTect is another method that uses Bayesian classifiers for detecting and annotating variants [34, 35]. MuTect has been widely used in cancer genomics research, especially in The Cancer Genome Atlas projects. Other VCF annotation tools are SnpEff [36] and SnpSift [37]. SnpEff can perform annotation for multiple variants and SnpSift allows rapid detection of significant variants from the VCF files [37]. The Variant Annotation Tool (VAT) distinguishes itself from other annotation tools in one aspect by adding cloud computing capabilities [38]. VAT annotation occurs at the transcript level to determine whether all or only a subset of the transcript isoforms of a gene is affected. VAT is dynamic in that it also annotates Multiple Nucleotide Polymorphisms (MNPs) and can be used on more than just the human species.

2.6. Database and Resources for Variant Filtration. During the annotation process, many resources and databases could be used as filtering criteria for detecting novel variants from common polymorphisms. These databases score a variant by its minor allelic frequency (MAF) within a specific population or study. The need for filtration of variants based on this number is subject to the purpose of the study. For example, Mendelian studies would be interested in including common SNVs while cancer studies usually focus on rare variants found in less than 1% of the population. NCBI dbSNP database, established in 2001, is an evolving database containing both well-known and rare variants from many organisms [39]. dbSNP also contains additional information including disease association, genotype origin, and somatic and germ-line variant information [39].

The Leiden Open Variation Database (LOVD) developed in 2005 links its database to several other repositories so that the user can make comparisons and gain further information [40]. One of the most popular SNV databases was developed in 2010 from the 1000 Genomes Project that uses statistics from the sequencing of more than 1000 “healthy” people of all ethnicities [41]. This is especially helpful for cancer studies, as damaging mutations found in cancer are often very rare in a healthy population. Another database essential for cancer studies is the Catalogue of Somatic Mutations In Cancer (COSMIC) [42]. This database of somatic mutations found in cancer studies from almost 20,000 publications allows for identification of potentially important cancer-related variants. More recently, the Exome Aggregation Consortium (ExAC) has assembled and reanalyzed WES data of 60,706 unrelated individuals from various disease-specific and population genetic studies [3]. The ExAC web portal and data provide a resource for assessing the significance of variants detected in WES data [3].

2.7. Functional Predictors of Mutation. Besides knowing if a particular variant has been previously identified, researchers may also want to determine the effect of a variant. Many functional prediction tools have been developed that all vary slightly in their algorithms. While individual prediction software can be used, ANNOVAR provides users with scores from several different functional predictors including SIFT, PolyPhen-2, LRT, FATHMM, MetaSVM, MetaLR, VEST, and CADD [34].

SIFT determines if a variant is deleterious using PSI-BLAST to determine conservation of amino acids based on closely related sequence alignments [43]. PolyPhen-2 uses a pipeline involving eight sequence based methods and three structure based methods in order to determine if a mutation is benign, probably deleterious, or known to be deleterious [44]. The Likelihood Ratio Test (LRT) uses conservation between closely related species to determine a mutations functional impact [45]. When three genomes underwent analysis
by SIFT, PolyPhen-2, and LRT, only 5% of all predicted deleterious mutations were agreed to be deleterious by all three methods [45]. Therefore, it has been shown that using multiple mutational predictors is necessary for detecting a wide range of deleterious SNVs. FATHMM employs sequence conservation within Hidden Markov Models for predicting the functional effects of protein missense mutations [46]. FATHMM weighs mutations based on their pathogenicity by the predicted interaction of the protein domain [46].

MetaSVM and MetaLR represent two ensemble methods that combine 10 predictor scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, and PhyloP) and the maximum frequency observed in the 1000 genomes populations for predicting the deleterious variants [47]. MetaSVM and MetaLR are based on the ensemble Support Vector Machine (SVM) and Logistic Regression (LR), respectively, for predicting the final variant scores [47].

The Variant Effect Scoring Tool (VEST) is similar to MetaSVM and MetaLR in that it uses a training set and machine learning to predict functionality of mutations [48]. The main difference in the VEST approach is that the training set and prediction methodology are specifically designed for Mendelian studies [48]. The Combined Annotation Dependent Depletion (CADD) method differentiates itself by integrating multiple variants with mutations that have survived natural selection as well as simulated mutations [49].

While all of these methods predict the functionality of a mutation, they all vary slightly in their methodological and biological assumptions. Dong et al. have recently tested the performance of these prediction algorithms on known datasets [47]. They pointed out that these methods rarely unanimously agree on if a mutation is deleterious. Therefore, it is important to consider the methodology of the predictor as well as the focus of the study when interpreting deleterious prediction results.

### 3. Computational Methods for Beyond VCF Analyses

After a VCF file has been generated, annotated, and filtered, there are several types of analyses that can be performed (Figure 2). Here we outline six major types of analyses that can be performed after the generation of a VCF file, with special focus on WES in cancer research: (i) significant somatic mutations, (ii) pathway analysis, (iii) copy number estimation, (iv) driver prediction, (v) linking variants to clinical information and actionable therapies, and (vi) emerging applications of WES in cancer research.

#### 3.1. Methods to Determine Significant Somatic Mutations

After VCF annotation, a WES sample can have thousands of SNVs identified; however, most of them will be silent (synonymous) mutations and will not be meaningful for follow-up study. Therefore, it is important to identify significant somatic mutations from these variants. Several tools have been developed to do this task for the analysis of cancer WES data, including SomaticSniper [50], MuTect [35], VarSim [51], and SomVarIUS [52].

SomaticSniper is a computational program that compares the normal and tumor samples to find out which mutations are unique to the tumor sample, hence predicted as somatic mutations [50]. SomaticSniper uses the genotype likelihood model of MAQ (as implemented in SAMtools) and then calculates the probability that the tumor and normal genotypes are different. The probability is reported as a somatic score which is the Phred-scaled probability. SomaticSniper has been applied in various cancer research studies to detect significant somatic variants.

Another popular somatic mutation identification tool is MuTect [35], developed by the Broad Institute. MuTect, like SomaticSniper, uses paired normal and cancer samples as input for detecting somatic mutations. After removing low-quality reads, MuTect uses a variant detection statistic to determine if a variant is more probable than a sequencing error. MuTect then searches for six types of known sequencing artifacts and removes them. A panel of normal samples as well as the dbSNP database is used for comparison to remove common polymorphisms. By doing this, the number of somatic mutations is not only identified but also reduced to a more probable set of candidate genes. MuTect has been widely used in Broad Institute cancer genomics studies.

While SomaticSniper and MuTect require data from both paired cancer and normal samples, VarSim [51] and SomVarIUS [52] do not require a normal sample to call somatic mutations. Unlike most programs of its kind, VarSim [51] uses a two-step process utilizing both simulation and experimental data for assessing alignment and variant calling accuracy. In the first step, VarSim simulates diploid genomes with germline and somatic mutations based on a realistic model that includes SNVs and SVs. In the second step, VarSim performs somatic variant detection using the simulated data and validates the cancer mutations in the tumor VCF. SomVarIUS is another recent computational method to detect somatic variants in cancer exomes without a normal paired sample [52]. In brief, SomVarIUS consists of 3 steps for somatic variant detection. SomVarIUS first prioritizes potential variant sites, estimates the probability of a sequencing error followed by the probability that an observed variant is germline or somatic. In samples with greater than 150x coverage, SomVarIUS identifies somatic variants with at least 67.7% precision and 64.6% recall rates, when compared with paired-tissue somatic variant calls in real tumor samples [52]. Both VarSim and SomVarIUS will be useful for cancer samples that lack the corresponding normal samples for somatic variant detection.

#### 3.2. Computational Tools for Estimating Copy Number Alteration

One active research area in WES data analysis is the development of computational methods for estimating copy number alterations (CNAs). Many tools have been developed for estimating CNAs from WES data based on paired normal-tumor samples such as CNV-seq [53], SegSeq [54], ADTEx [55], CONTRA [56], EXCAVATOR [57], ExomeCNV [58], Control-FREEC (control-FREE Copy number caller) [59], and CNVseequer [60]. For example, VarScan2 [61] is a computational tool that can estimate somatic mutations and CNAs from paired normal-tumor samples. VarScan2 utilizes...
a normal sample to find Somatic CNAs (SCNAs) by first comparing Q20 read depths between normal and tumor samples and normalizes them based on the amount of input data for each sample [61]. Copy number alteration is inferred from the log2 of the ratio of tumor depth to normal depth for each region [61]. Lastly, the circular segmentation (CBS) algorithm [62] is utilized to merge adjacent segments to call a set of SCNAs. These SCNAs could be further classified as large-scale (>25% of chromosome arm) or focal (<25%) events in the WES data [63].

Recently, ExomeAI was developed to detect Allelic Imbalance (AI) from WES data [64]. Utilizing heterozygous sites, ExomeAI finds deviations from the expected 1:1 ratio between an A- and B-allele in multiple tumor samples without a normal comparison. Absolute deviation of B-allele frequency from .05 is calculated and similar to VarScan2; the CBS algorithm is applied to each chromosomal arm [62]. In order to reduce the number of false positives, a database was created with 500 (and counting) normal samples to filter out known AIs. This represents a novel tool to analyze WES for the detection of recurrent AI events without matched normal samples.

A systematic evaluation of somatic copy number estimation tools for WES data has been recently published [63]. In this study, six computational tools for CNAs detection (AD Tex, CONTRA, Control-FREC, EXCAVATOR, ExomeCNV, and VarScan2) were evaluated using WES data from three TCGA datasets. Using a SNP array as the reference, this study found that these algorithms gave highly variable results. The authors found that AD Tex and EXCAVATOR had the best performance with relatively high precision and sensitivity when compared to the reference set. The study showed that the current CNA detection tools for WES data still have limitations and called for more robust algorithms for this challenging task.

3.3. Computational Tools for Predicting Drivers in Cancer Exomes. Cancer is a disease driven by genetic variations and copy number alterations. These genetic events can be classified into two classes, “driver” and “passenger” mutations. Driver mutations are the key mutation that drive the development of cancer and provide a survival advantage, whereas passenger mutations are “by-stander” alterations that happen to be altered in the primary cells but do not provide a survival advantage. As the cancer exomes tend to have high mutational burdens, identifying the “driver” mutations from the “passenger” mutations is one of the key analyses in cancer research. Several tools have been developed to find driver mutations including but not limited to CHASM [65], Dendrix [66], and MutSigCV [67].

CHASM (Cancer-specific High-throughput Annotation of Somatic Mutations) uses random forest as the machine learning approach to distinguish the difference between driver and passenger mutations in cancer [65]. CHASM was trained on the curated driver mutations obtained from the COSMIC database (“positive examples”) and synthetic passenger mutations generated according to the background of base substitution frequencies observed in specific tumor types (“negative examples”). CHASM can achieve high sensitivity and specificity when discriminating between known driver missense mutations and randomly generated missense mutations when tested in real tumor samples. This method has been one of the popular driver detection prediction tools for cancer researchers and has been applied in various cancer genomic studies.

Another common driver mutation tool is MutSigCV developed to resolve the problem of extensive false-positive findings that overshadow true driver mutations [67]. As the size of cancer genomes sequenced has increased implausible genes (such as TTN) have been falsely reported as being related to cancer when in fact their large size just makes the probability they would be mutated by chance increase [67]. MutSigCV takes into account patient-specific mutation frequency and spectrum as well as gene-specific background mutation rates, expression level, and replication time. By pooling all of this available data into one tool, MutSigCV has become a standard tool used for driver mutation identification in cancer studies.

De novo Driver Exclusivity (Dendrix) is a novel computational tool to determine de novo driver pathways (gene sets) from somatic mutations in patient data [66]. The main goal of the Dendrix algorithm is to find gene sets with high coverage and high exclusivity properties from the somatic data. The high coverage property assumes most patients have at least one driver mutation in the gene set, whereas the high exclusivity property assumes that these driver mutations are rarely mutated together in the same patient. Two algorithms were developed in Dendrix, one based on a greedy algorithm and one based on the Markov Chain Monte Carlo (MCMC) algorithm, to measure sets of genes that exhibit both criteria. When Dendrix was applied to the TCGA data, the algorithms identified groups of genes that were mutated in large subsets of patients and these mutations were mutually exclusive. This tool provides an opportunity to analyze WES data to identify driver pathways in cancer genomic studies.

3.4. Methods for Pathway Analysis. After candidate somatic mutations have been identified; one common type of analysis is to determine which pathways are affected by these mutations. Common pathway resources and tools used for these types of analysis include KEGG [68], DAVID [69], STRING [70], BEReX [71], DAPPLE [72], and SNPsea [73].

KEGG represents one of the most popular databases for pathway analysis. DAVID is a popular online tool for performing functional enrichment analysis based on user defined gene lists. STRING is the largest protein-protein interactions database for querying and searching for interactions between user defined gene lists. BEReX integrates STRING, KEGG, and other data sources to explore biomedical interactions between genes, drugs, pathways, and diseases. Both STRING and BEReX allow users to perform functional enrichment analysis and the flexibility to explore the interactions between user defined gene lists by expanding the networks.

DAPPLE (Disease Association Protein-Protein Link Evaluator) uses literature reported protein-protein interactions to identify significant physical connectivity among the
genes of interest [72]. DAPPLE hypothesizes that genetic variation affects underlying mechanisms only detectable by protein-protein interactions [72]. SNPsea is another pathway analysis tool that requires specific SNP data [73]. SNPsea calculates linkage disequilibrium between involved genes and uses a sampling approach to determine conditions that are affected by these interactions.

3.5. Computational Tools for Linking Variants to Treatments.

The ability to link variants with actionable drug targets is an emerging research topic in precision medicine. Databases such as My Cancer Genome have provided the framework for these studies (https://www.mycancergenome.org/). My Cancer Genome provides a bridge between genomic data and clinical therapeutic treatments. Similarly, ClinVar provides information on the relationship between variants and clinical therapy [74]. By collecting both the variants and the clinical significance related to these variants, ClinVar offers a database for researchers to explore the significance of sequencing findings in the clinical setting [74]. Pharmacological databases such as PharmGKB [75], DrugBank [76], and DSigDB [77] provide the link between drug and drug targets (variants). For example, by querying a list of variants to one of these databases, it allows users to identify actionable targets via enrichment analysis for the repurposing of drugs.

Similarly, the ability to incorporate clinical data into sequencing studies is vital to the advancement of personalized medicine. However, due to the lack of integration between electronic health records (EHR) and molecular analysis, this remains one of the challenges in translating WES data analysis into clinical practice. Projects such as cBioPortal provide a framework for incorporating sequencing data with available clinical data [78]. New methods for addressing this task are urgently needed to take advantage of the important applications of WES data within the clinic in order to advance precision medicine.

4. WES Analysis Pipelines

WES data analysis pipelines integrate computational tools and methods described in the previous sections in a single analysis workflow. Here, we review three recent sequencing pipelines SeqMule [79], Fastq2vcf [80], and IMPACT [81] that assimilate some of the tools described in previous sections.

SeqMule stands out in part due to the use of five alignment tools (BWA, Bowtie 1 and 2, SOAP2, and SNAP) and five different variant calling algorithms (GATK, SAMtools, VarScan2, FreeBayes, and SOAPsnp) [79]. SeqMule contains at least one feature that performs Pre-VCF analyses to generate a filtered VCF file. SeqMule also generates an accompanying HTML-based report and images to show an overview of every step in the pipeline. Fastq2vcf also performs the Pre-VCF analyses using BWA as an alignment tool and variant calling by GATK, UnifiedGenotyper, HaplotypeCaller, SAMtools, and SNVer resulting in a filtered VCF after implementation of ANNOVAR and VEP [80]. Fastq2vcf can be used in a single or parallel computing environment on variety of sequencing data.

Both SeqMule and Fastq2vcf pipelines focus on taking raw sequencing data and converting it into a filtered VCF file. IMPACT (Integrating Molecular Profiles with ACtionable Therapeutics) WES data analysis pipeline was developed to take this analysis a step further by linking a filtered VCF to actionable therapeutics [81]. The IMPACT pipeline contains four analytical modules: detecting somatic variants; calling copy number alterations; predicting drugs against the deleterious variants; and tumor heterogeneity analysis. IMPACT has been applied to longitudinal samples obtained from a melanoma patient and identified novel acquired resistance mutations to treatment. IMPACT analysis revealed loss of CDKN2A as a novel resistance mechanism to the combination of dabrafenib and trametinib treatment and predicted potential drugs for further pharmacological and biological studies [81].

To compare the strengths and weaknesses between these three WES pipelines, SeqMule allows the use of different alignment algorithms in its pipeline whereas IMPACT and Fastq2vcf only utilize BWA as the sequencing alignment algorithm. SAMtools is the common tool used by IMPACT, Fastq2vcf, and SeqMule to call variants. In addition, Fastq2vcf and SeqMule employ GATK and other variant calling algorithms for variants detection. Fastq2vcf and IMPACT both annotate the variants with ANNOVAR. Fastq2vcf also utilizes VEP and IMPACT utilizes SIFT and PolyPhen-2 as the primary variants prediction methods. For Post-VCF analysis, IMPACT pipeline has more options as compared to SeqMule and Fastq2vcf. In particular, IMPACT performs copy number analysis, tumor heterogeneity, and linking of actionable therapeutics to the molecular profiles. However, IMPACT is only designed to be performed on tumor samples while SeqMule and Fastq2vcf are designed for any WES dataset. Therefore, it is advisable for the users to consider the analytic needs to select the appropriate WES data analysis pipeline for their research.

As recently discussed by Altman et al., part of the U.S. Precision Medicine Initiative (PMI) includes being able to define a gold standard of pipelines and tools for specific sequencing studies to enable a new era of medicine [82]. Automated pipelines such as these will accelerate the analysis and interpretation of WES data. Future development of data analysis pipeline will be needed to incorporate newer and wider tools tailored for specific research questions.

5. Conclusions

In summary, we have reviewed several computational tools for the analysis and interpretation of WES data. These computational methods were developed to generate VCF files from raw sequencing data, as well as tools that perform downstream analyses in WES studies. Each tool has specific strengths and weaknesses, and it appears that using several of them in combination would lead to more accurate results. Currently, there are still challenges for bioinformaticians at every step in analyzing WES data. However, the greatest area of need is in the development of tools that can link the information found in a VCF file to clinical databases and therapeutics. Research in this area will help to advance
precision medicine by providing user-friendly and informative knowledge to transcend the laboratory.

**Competing Interests**

The authors declare that there are no competing interests regarding the publication of this paper.

**Acknowledgments**

The authors would like to thank the Translational Bioinformatics and Cancer Systems Biology Lab members for their constructive comments on this manuscript. This work is partly supported by the National Institutes of Health P50CA058187, Cancer League of Colorado, the David F. and Margaret T. Grohne Family Foundation, the Rifkin Endowed Chair (WAR), and the Moore Family Foundation.

**References**


Research Article

GPA-MDS: A Visualization Approach to Investigate Genetic Architecture among Phenotypes Using GWAS Results

Wei Wei,1 Paula S. Ramos,1,2 Kelly J. Hunt,1,3 Bethany J. Wolf,1 Gary Hardiman,1,2,4 and Dongjun Chung1

1Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA
2Department of Medicine, Medical University of South Carolina, Charleston, SC, USA
3Ralph H. Johnson VA Medical Center, Charleston, SC, USA
4Center for Genomic Medicine, Medical University of South Carolina, Charleston, SC, USA

Correspondence should be addressed to Dongjun Chung; chungd@musc.edu

Received 27 May 2016; Accepted 5 October 2016

Academic Editor: Brian Wigdahl

Copyright © 2016 Wei Wei et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Genome-wide association studies (GWAS) have identified tens of thousands of genetic variants associated with hundreds of phenotypes and diseases, which have provided clinical and medical benefits to patients with novel biomarkers and therapeutic targets. Recently, there has been accumulating evidence suggesting that different complex traits share a common risk basis, namely, pleiotropy. Previously, a statistical method, namely, GPA (Genetic analysis incorporating Pleiotropy and Annotation), was developed to improve identification of risk variants and to investigate pleiotropic structure through a joint analysis of multiple GWAS datasets. While GPA provides a statistically rigorous framework to evaluate pleiotropy between phenotypes, it is still not trivial to investigate genetic relationships among a large number of phenotypes using the GPA framework. In order to address this challenge, in this paper, we propose a novel approach, GPA-MDS, to visualize genetic relationships among phenotypes using the GPA algorithm and multidimensional scaling (MDS). This tool will help researchers to investigate common etiology among diseases, which can potentially lead to development of common treatments across diseases. We evaluate the proposed GPA-MDS framework using a simulation study and apply it to jointly analyze GWAS datasets examining 18 unique phenotypes, which helps reveal the shared genetic architecture of these phenotypes.

1. Introduction

Genome-wide association studies (GWAS) have been conducted to study the genetic basis of complex human traits. As of August 2015, more than 15,000 single nucleotide polymorphisms (SNPs) have been reported to be significantly associated with at least one complex trait (the NHGRI-EBI catalog of published GWAS [1], https://www.ebi.ac.uk/gwas/). “Pleiotropy,” that is, the sharing of genetic factors among complex traits, is well documented and a systematic analysis of the GWAS catalog of published GWAS showed that 17% of the reported genes are associated with multiple traits [2]. For example, genetic studies for five psychiatric disorders suggested a very strong genetic correlation between schizophrenia and bipolar disorder [3, 4]. Pleiotropy has also been demonstrated among several other types of traits, such as cancers [5].

In order to leverage pleiotropy between complex traits and effectively integrate multiple GWAS datasets, Chung et al. [6] developed a unified statistical framework, named GPA (Genetic analysis incorporating Pleiotropy and Annotation), which provides statistically rigorous and biologically interpretable inference tools for genetic studies. Application of GPA to five psychiatric disorder GWAS datasets from the Psychiatric Genomics Consortium [3, 4] showed that GPA can accurately identify pleiotropic structure among these diseases [6]. While the GPA framework provides a statistically rigorous framework to evaluate pleiotropy, it still remains limited to a small number of phenotypes and it is common to consider a joint analysis of only two phenotypes mainly
due to computational efficiency and estimation stability and robustness. In practice, we are interested in jointly studying larger numbers of phenotypes; however it is not a trivial task to investigate and integrate results from multiple pairs of phenotypes.

In order to address this challenge, in this paper, we propose a novel visualization approach, GPA-MDS, to investigate genetic architecture with a joint analysis of multiple GWAS datasets using the GPA algorithm and the multidimensional scaling (MDS) approach. Specifically, the GPA algorithm allows for evaluation of pleiotropy between two phenotypes within a statistically rigorous framework. Then, the MDS approach effectively integrates these results for a large number of phenotypes and provides a two-dimensional map of genetic architecture. This paper is organized as follows. In Section 2, we review the GPA and MDS algorithms and propose the GPA-MDS approach. In Section 3, we evaluate the proposed method with a simulation study and apply it to a joint analysis of 18 GWAS datasets. Finally, in Section 4, we will discuss future research directions.

2. Methods: GPA-MDS Approach

Figure 1 shows a diagram of overall workflow for the GPA-MDS framework. Here, by taking association $p$-value for each SNP from each GWAS study as an input, we first calculate a distance matrix between phenotypes using the pleiotropy hypothesis testing procedures in the GPA framework. Then, we generate a plot depicting a global picture of genetic relationship among phenotypes by projecting phenotypes onto two-dimensional space using the multidimensional scaling (MDS) algorithm based on this distance matrix.

2.1. Statistical Inference of Pleiotropy Using the GPA Algorithm. In this section, we review the GPA framework [6] for the case of a joint analysis of two GWAS datasets. Although there is no limitation in the number of GWAS datasets that can be jointly analyzed in the GPA framework, a joint analysis of two GWAS datasets is often appropriate in the sense of the computational efficiency and estimation stability and robustness. Let $t$ be the index for SNPs and let $k$ be the index for GWAS datasets. Suppose that we have performed hypothesis testing of genome-wide SNPs for two GWAS and obtained their $p$-values. Specifically, for GWAS1, we have

$$ H_0^{(M)}:$$

$$H_0^{(M)}:$$

$$p$$-value for GWAS1: $P_{11}, P_{21}, \ldots, P_{M1}$,

where $M$ is the number of SNPs and $P_{kl}$ denotes $p$-value of the $l$th SNP in the $k$th GWAS. Similarly, for GWAS2, we have

$$H_0^{(M_2)},$$

$$p$$-value for GWAS2: $P_{12}, P_{22}, \ldots, P_{M2}$.

Let us denote $P_1 = (P_{11}, P_{21}, \ldots, P_{M1})$ and $P_2 = (P_{12}, P_{22}, \ldots, P_{M2})$.

We introduce latent variables $Z_t = [Z_{00}, Z_{01}, Z_{01}, Z_{11}]$ indicating the association between the $t$th SNP and the two phenotypes: $Z_{00} = 1$ means that the $t$th SNP is not associated with any phenotypes, $Z_{10} = 1$ means that it is only associated with the first one, $Z_{01} = 1$ means that it is only associated with the second one, and $Z_{11} = 1$ means that it is associated with both. We assume that $Z_{00}, Z_{10}, Z_{01}, Z_{11} \in \{0, 1\}$ and $Z_{00} + Z_{10} + Z_{01} + Z_{11} = 1$ because a SNP can only be
one of these states. Given these latent variables, we assume the following emission distributions:

\[ \begin{align*}
\pi_{00} &= \Pr(Z_{000} = 1): \\
(P_{11} \mid Z_{000} = 1) &\sim \mathcal{U}[0, 1], (P_{12} \mid Z_{000} = 1) &\sim \mathcal{U}[0, 1], \\
\pi_{01} &= \Pr(Z_{100} = 1): \\
(P_{21} \mid Z_{100} = 1) &\sim \text{Beta}(\alpha_1, 1), (P_{22} \mid Z_{100} = 1) &\sim \mathcal{U}[0, 1], \\
\pi_{11} &= \Pr(Z_{200} = 1): \\
(P_{31} \mid Z_{200} = 1) &\sim \mathcal{U}[0, 1], (P_{32} \mid Z_{200} = 1) &\sim \text{Beta}(\alpha_2, 1), \\
\end{align*} \]

where \( 0 < \alpha_k < 1, k = 1, 2 \). We put the constraint \( 0 < \alpha_k < 1 \) to model that a smaller \( p \)-value is more likely than a larger \( p \)-value when it is from the nonnull group [7]. Parameters in the GPA model can be estimated using the Expectation-Maximization (EM) algorithm [8], which is remarkably computationally efficient because we have explicit formulas with the following hypothesis:

\[ \begin{align*}
H_0: \pi_{11} &= \pi_{1*} + \pi_{*1}, \text{ versus } H_1: \text{not } H_0,
\end{align*} \]

where \( \pi_{1*} = \pi_{10} + \pi_{11} \) and \( \pi_{*1} = \pi_{01} + \pi_{11} \). The likelihood ratio test (LRT) statistic can be constructed as follows:

\[ \lambda = \frac{\Pr \left( \left( P_1, P_2; \widehat{\Theta}_0 \right) \right)}{\Pr \left( \left( P_1, P_2; \Theta \right) \right)}, \]

where \( \widehat{\Theta}_0 \) represents the parameter estimates obtained under the null hypothesis of pleiotropy test. The test statistic \((-2 \log \lambda\) asymptotically follows \( \chi^2 \)-distribution with degree of freedom of one, under the null hypothesis. Fitting the GPA model and hypothesis testing of pleiotropy were implemented as a part of the R package “GPA,” which is currently available in its GitHub page (http://dongjunchung.github.io/GPA/).

2.2. Visualization of Pleiotropic Structure Using Multidimensional Scaling. For the visualization of genetic relationships among phenotypes, we first run pleiotropy tests for all possible pairs of GWAS datasets and generate a matrix of their \( \log_{10} \)-transformed \( p \)-values, denoted as \( s_{ij} \). Then, we define a distance between \( i \)th and \( j \)th phenotypes as \( d_{ij} = s_{ij} - 2 \min_{k \in \{1, \ldots, n\}}(s_{ik}) \). Note that this definition of distance assigns a shorter distance to a pair of phenotypes with smaller pleiotropy test \( p \)-values, while it also allows avoiding negative distance values. Then, we feed this distance matrix to the MDS algorithm and project phenotypes onto the two-dimensional space. As a result, MDS essentially clusters phenotypes based on their genetic similarities; that is, phenotypes sharing fewer SNPs are located further apart on the two-dimensional space compared to phenotypes sharing more SNPs. Specifically, given a distance matrix \( D = (d_{ij}) \), MDS seeks to find \( x_1, \ldots, x_n \in \mathbb{R}^p \) such that

\[ d_{ij} = \|x_i - x_j\|_2 \]

by minimizing the following objective function:

\[ \sum_{i \neq j} (d_{ij} - d_{ij})^2. \]

Here, we consider \( p = 2 \) to provide easily understandable visualization. We used the function `cmdscale()` in R with default settings to implement MDS.

3. Results

3.1. Simulation Study. We conducted a simulation study to evaluate the performance of GPA-MDS approach. Here, we assume that there are five GWAS datasets, each of which is profiled for a set of 10,000 SNPs common to all 5 datasets. Among these 10,000 SNPs, 2,000 SNPs (20%) were assumed to be risk SNPs for each phenotype. In order to generate pleiotropic structure, we set 75% of the risk SNPs to be shared between phenotypes 1 and 2 and also between phenotypes 3 and 4, while no risk SNPs are shared between phenotypes 1/2 and phenotypes 3/4. Phenotype 5 did not share any risk SNPs with any other phenotypes as a negative control.
Figure 2: Design of simulation study. Each box indicates a GWAS study and x-axis represents the SNP index. The gray box within each box indicates risk SNPs. In this study, we considered five phenotypes and 20% of the SNPs were assumed to be risk SNPs for each phenotype. We further assumed that 75% of risk SNPs were shared between phenotypes 1 and 2 and also between phenotypes 3 and 4 to generate pleiotropic structure. Phenotype 5 did not share any risk SNPs with any other phenotypes as a negative control.

Finally, for each phenotype, we simulated $p$-values for nonrisk SNPs from a uniform distribution and $p$-values for risk SNPs from a Beta distribution with parameters 0.4 and 1. Figure 3 shows the GPA-MDS plot for these five phenotypes. In this plot, phenotypes 1 and 2 are clustered and phenotypes 3 and 4 generate another cluster. Phenotype 5 is isolated and located away from these two phenotype clusters. This result shows that the proposed GPA-MDS approach can provide easily interpretable visualization revealing the pleiotropic architecture among phenotypes.

3.2. Real Data Analysis. We applied the proposed GPA-MDS approach to the GWAS datasets of 18 phenotypes, using summary statistics, which are publicly available from consortium websites. Specifically, we considered (1) attention deficit/hyperactivity disorder (ADHD), autism spectrum disorder (ASD), bipolar disorder (BPD), major depressive disorder (MDD), and schizophrenia (SCZ) from the Psychiatric Genomics Consortium (http://www.med.unc.edu/pgc); (2) Crohn's disease (CD) and ulcerative colitis (UC) from the International Inflammatory Bowel Disease Genetics Consortium (https://www.ibdgenetics.org/); (3) rheumatoid arthritis (RA) (https://www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl.etal.2010NG/); (4) high-density lipoprotein (HDL), low-density lipoprotein (HDL), triglycerides (TG), and total cholesterol (TC) from the Global Lipids Consortium (http://csg.sph.umich.edu/abecasis/public/lipids2010/); (5) type 2 diabetes (T2D) from the DIAbetes Genetcs Replication And Meta-analysis Consortium (http://diagram-consortium.org/); (6) coronary artery disease (CAD) from the CARDIoGRAM Consortium (http://www.cardiogramplusc4d.org/data-downloads/); (7) systolic blood pressure (SBP) and diastolic blood pressure (DBP) from the International Consortium for Blood Pressure (http://www.georgehretnlab.org/icbp_088023401234-9812599.html); and (8) fasting glucose (FG) and log of fasting insulin (LFI) from the MAGiC Consortium. We used the intersection of SNPs among these datasets, which consists of 228,944 SNPs.

Figure 4 shows the GPA-MDS plot for the 18 phenotypes. We can see that clinically related phenotypes are tightly
clustered in this plot. For example, all the neuropsychiatric disorders (ADHD, ASD, BPD, MDD, and SCZ) generate a cluster, inflammatory bowel diseases (UC and CD) make a cluster, lipid-related phenotypes (HDL, LDL, TC, and TG) cluster together, blood pressure phenotypes (SBP and DBP) cluster, and so on. Moreover, RA is also located relatively close to UC and CD, which is consistent with the literature as RA, UC, and CD are all autoimmune diseases [11]. The cluster containing both T2D and CAD in this plot is also well supported by prior studies, which suggest the pleiotropy between T2D and CAD [12–14]. These results show the potential of the proposed GPA-MDS approach for the investigation of pleiotropic architecture, which can be used to promote understanding of common etiology and development of joint treatment of diseases.

In order to further understand the phenotype mapping provided by GPA-MDS, we checked the number of risk SNPs shared among phenotypes (Figure 5). Here, “risk SNPs” were determined using the GPA algorithm by controlling the global FDR at 0.1. We can see that some of the phenotypes that are closely located in the GPA-MDS plot actually share more risk SNPs, as in the case of CD-UC, LDL-TG-TG, and SBP-DBP. However, it might look like that Figure 5 seems to contradict the GPA-MDS plot for other phenotypes. For example, although ADHD and ASD are located relatively close to each other in the GPA-MDS plot, it seems that only a few risk SNPs are shared between these two phenotypes. Such “discrepancy” happens because GPA evaluates pleiotropy by checking whether the number of shared risk SNPs ($\tau_{ij}$) is significantly higher than what is expected by chance ($\tau_{ij1}$), not simply based on the counts of shared risk SNPs ($\tau_{ij}$).

In order to confirm this explanation and further evaluate the utility of employing the GPA algorithm, we generated a MDS plot of phenotypes, where the distance between two phenotypes was determined solely by the number of shared risk SNPs (Figure 6). Specifically, we generated the MDS plot by defining the distance between $i$th and $j$th phenotypes as $d_{ij} = 2 \max_{k,l} |n_{ij1} - n_{ij2}|$, where $n_{ij}$ is the log$_{10}$-transformed risk SNPs shared between $i$th and $j$th phenotypes. In this mapping of phenotypes, clinically related phenotypes failed to cluster together but instead multiple phenotype groups are mixed together. Furthermore, we can see that the mapping is essentially driven by a few pairs of phenotypes which share large numbers of genotypes, such as CD-UC, TC-LDL, and SBP-DBP. Hence, we can conclude that it was actually critical to utilize the GPA algorithm to evaluate pleiotropy in the first step of our visualization framework because it provides biologically more meaningful visualization of genetic relationship among phenotypes.

4. Conclusion

In this paper, we proposed a novel visualization approach for genetic architecture among phenotypes, namely, GPA-MDS, using the GPA algorithm and the multidimensional scaling (MDS) approach. While the GPA framework provides a rigorous evaluation of pleiotropy between a pair of phenotypes, the MDS approach extends this investigation to larger number of phenotypes in a computationally efficient way. The application of GPA-MDS to the genetic studies of 18 phenotypes revealed patterns of shared genetic architecture among phenotypes, underscoring the potential of
the proposed method to investigate genetic sharing among complex traits. We note that when the proposed GPA-MDS framework is used, it is critical to confirm that its assumptions hold well for the input GWAS dataset. Specifically, because GPA-MDS uses GPA as its first step, users need to confirm that GWAS association \( p \)-values provided to GPA-MDS satisfy the GPA assumptions, for example, uniformity of null \( p \)-values. For example, if population stratification and cryptic relatedness have not been accounted for in previous GWAS studies, \( p \)-values of null SNPs may not follow the assumed uniform distribution and can result in biased MDS visualization results. Hence, these confounding effects should be checked carefully and addressed before applying the GPA-MDS framework to these GWAS association \( p \)-values. We recommend readers to check [6] for deeper discussion of the GPA assumptions.

Currently, we are working on the following directions that can further improve the GPA-MDS framework. First, in this paper, we used a definition of distance based on the logarithm transformation of \( p \)-values. While this approach is intuitive and works well in practice, other choices of distance measures might change visualization results. Hence, it is of great interest to investigate other choices of distance measures and their impacts on visualization results. Second, while the proposed GPA-MDS approach promotes global understanding of genetic architecture, it is still laborious to pinpoint risk SNPs leading to phenotype clusters when we investigate a large number of phenotypes. Hence, it would be desirable to automate the procedure to identify overlapping risk SNPs. We expect that GPA-MDS will be a useful method in elucidating the pleiotropic architecture of complex traits, which can contribute to a better understanding of shared genetic mechanisms and the development of improved diagnosis and therapeutics.

**Disclosure**

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**Competing Interests**

The authors declare that there are no competing interests regarding the publication of this paper.

**Acknowledgments**

This work was supported by the US National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) grants K01 AR067280 (Paula S. Ramos) and P60 AR062755 (Paula S. Ramos and Bethany J. Wolf), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) grant R01 DK088778 (Kelly J. Hunt), National Institute of General Medical Sciences (NIGMS) grant R01 GM122078, and National Cancer Institute (NCI) grant R21 CA209848 (Dongjun Chung and Gary Hardiman) from the National Institutes of Health (NIH).

**References**


Embracing Integrative Multiomics Approaches

Daniel M. Rotroff and Alison A. Motsinger-Reif

1Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27607, USA
2Department of Statistics, North Carolina State University, Raleigh, NC 27607, USA

Correspondence should be addressed to Alison A. Motsinger-Reif; alison.motsinger@gmail.com

Received 13 May 2016; Accepted 22 June 2016

Copyright © 2016 D. M. Rotroff and A. A. Motsinger-Reif. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As “-omics” data technology advances and becomes more readily accessible to address complex biological questions, increasing amount of cross “-omics” dataset is inspiring the use and development of integrative bioinformatics analysis. In the current review, we discuss multiple options for integrating data across “-omes” for a range of study designs. We discuss established methods for such analysis and point the reader to in-depth discussions for the various topics. Additionally, we discuss challenges and new directions in the area.

1. Introduction

The past decade has witnessed tremendous advancements in biotechnology and computational performance that have provided vast amounts of new data and accompanying optimism for the burgeoning improvements to human health and disease treatment. It is now possible, and increasingly routine, for studies to test thousands to millions of molecular endpoints. However, as the dimensionality of these data increases, larger sample sizes are required, and studies are now being conducted on an unprecedented scale. We are now seeing an explosion of data in almost every area of disease and clinical research. These technologies are commonly referred to as “omics” technologies, related to the suffix “-ome,” defined as “all constituents considered collectively” [1]. We now have vast amounts of data related to the genome, transcriptome, epigenome, proteome, and metabolome. In fact, many of these areas of research have spawned subfields that are rapidly advancing our mechanistic understanding of biology (e.g., pharmacogenomics, metagenomics, lipidomics, kinomics, and secretomics). However, too often we as researchers find small amounts of variation explained and are left with “missing heritability” and unexplained variation, reinforcing the seemingly exponential complexity of biology [2]. It seems as Robert M. Persig said that “the number of rational hypotheses that can explain any given phenomenon is infinite” [3]. Although it is clear that no single “-omics” technology can fully capture the intricacy of most complex diseases or other clinically relevant traits, the collective information from each of these technology platforms when combined has the potential to offer incredible insight into the mechanisms of complex disease and other important clinical traits.

Although it is clear that data integration is required, methods for achieving this are far from systematic. The integrative genomics methodologies that are used to interpret these data require expertise in multiple different disciplines, such as biology, medicine, mathematics, statistics, and bioinformatics. Such interdisciplinary approaches require diverse expertise, either through extensive interdisciplinary training or through extensive collaborations. The accumulation of enormous quantities of molecular data has led to the emergence of “systems biology”—a branch of science that discovers the principles that underlie the basic functional properties of living organisms, starting from interactions between macromolecules. Integrative genomics is based on the fundamental principle that any biological mechanism builds upon multiple molecular phenomena, and only through the understanding of the interplay within and between different layers of genomic structures can one attempt to fully understand phenotypic traits. Therefore, principles of integrative genomics are based on the study of...
molecular events at different levels and on the attempt to integrate their effects in a functional or causal framework.

2. Tools for Integrative Analysis

2.1. Using Publically Available Databases. Commonly used approaches involve linking all markers at the genomic, proteomic, metabolomics, and other levels back to annotated genes. In general, this approach works sufficiently because well annotated and curated databases describing genes and their known biological functions are readily available, though the various sources of data can be a challenge for analysis. Examples of these databases include NCBI’s gene database (http://www.ncbi.nlm.nih.gov/gene/), gene ontology (GO) (http://geneontology.org/), Ensembl (http://useast.ensembl.org), KEGG (http://www.genome.jp/kegg/pathway.html), HMDB (http://www.hmdb.ca/), MetaCyc (http://metacyc.org/), WikiPathways (http://www.wikipathways.org/index.php/WikiPathways), and DAVID (http://david.abcc.ncifcrf.gov/), and many others are also available. For data that is more granular than the “gene level” (e.g., SNPs, CpGs), methods for combining dependent univariate test statistics or \( p \) values are now available (e.g., SKAT [4], Correlated Lancaster Approach [5], and decorrelation tests [6]). As an example, the Correlated Lancaster Approach is a modified version of the Fisher method for combining multiple \( p \) values; however, when \( p \) values are correlated the Fisher method for combining \( p \) values will cause inflation of Type I error rates [5]. The Correlated Lancaster Approach addresses this by accounting for the underlying correlation structure of \( p \) values to limit Type I error and allowing for \( p \) values from multiple tests to be aggregated appropriately [5]. Now that resources such as 1000 genomes (http://www.1000genomes.org/) are available, methods for genotype imputation [7, 8] have made it possible to merge different genotyping platforms therefore greatly enhancing the ability to integrate genomics data and perform meta-analyses.

However, some data types are not readily mapped to annotated genes and these annotation limitations are particularly noticeable for the newest “omics” technologies. Metabolomics, for example, has major gaps in annotation that limit integration potential and limit the utility of pathway based and integrative methods approaches [9]. Metabolomics data is typically interpreted in the context of metabolic pathways and KEGG is an example of a database that contains metabolic pathways consisting of both metabolites and enzymes organized into groups related to metabolism, cellular processes, human diseases, and others. However, the lack of annotated metabolites indicates that we still have much to learn about the role of many metabolites in human health. Improved understanding of how genetic variants affect downstream molecular changes, such as metabolite levels, will be critical to improving our ability to interpret and integrate these types of data.

Once the results are mapped to annotations in a database, various integrative analysis approaches can be taken. While “integrative analysis” and “systems biology” can be vaguely defined workflows, in the current discussion we will consider the analysis of at least two different types of omics data as integrative.

The analysis can be restricted to molecular data (such as in expression quantitative trait loci (eQTL) studies, in which the relation between germ line variation and gene expression is investigated) or it can involve clinical outcomes (e.g., disease status or treatment response) or intermediate phenotypes and biomarkers.

2.2. Selecting the Appropriate Analysis Strategy. While it is possible to design an analysis plan to ask a variety of interesting biological and clinical questions, there are a few themes that usually emerge. The first common objective of analysis is to understand molecular behaviors, mechanisms, and relationships between and within the different types of molecular structures, including associations between these and various phenotypes, such as clinical outcomes and pathways. The second objective is often to understand the taxonomy of diseases or other clinical traits, thereby classifying individuals into latent classes of disease subtype; and the third objective is to predict an outcome or phenotype for prospective patients. Some statistical methods are specialized to one type of question, and others can be used for several. Some of the tools, such as enrichment analysis, were originally designed to reveal features of genes and pathways, whereas others, such as integrative clustering, were designed to reveal features of patient subgroups; however, most of the tools discussed below can be applied to both. The statistical methods used can be unsupervised or supervised (e.g., according to whether one proceeds in an exploratory manner or applies clinical labels to individual cases). Often these methods are used in conjunction with cross-validation or other model selection approaches to prevent overfitting.

2.3. Sequential Analysis. One of the most commonly used approaches, because of its ready application and interpretation, is sequential analysis. In sequential analysis, evidences (measures of association, etc.) from distinct omics levels are used. This approach allows the confirmation or refinement of findings based on one data type, with additional analyses of further omics data obtained from the same set of samples. In this case, at least two types of omics data are analyzed, for example, copy-number variants (CNVs) and gene expression level data.

Typically, in sequential approaches, an analysis of each dataset is made independently of the others and produces a list of interesting entities (omics level variables), which are then linked to each other. For example, differentially expressed genes in one list are compared with each other and then with different CNVs that have been matched to the closest gene in a second list. Usually, the lists are intersected to find the genes that are confirmed in the analysis of each data type. Comparing ranks of each gene in each list leads to measures of concurrence. If each entity in each list has a value of association with the outcome of interest (e.g., a \( t \)-test statistic) then these values can be combined, though there are challenges in how to create a combined \( p \) value to this intersection after proper controls for multiple comparison.
Approaches for combining $p$ values and permutation testing are suggested approaches. Occasionally, the various analyses are not performed in parallel but as a sequence of filtering steps, each functioning on a single data type. This approach can simplify statistical inference, but the results are highly dependent on ordering of the steps in the sequence. Such differences can be difficult to interpret and add a layer of complexity to ensuring reproducibility of analysis. Because of this, details of methods, including annotation details, must be shared at the level of databases and code to ensure reproducibility.

2.4. Gene Set and Pathway Based Analyses. Another very important area of integrative analysis is gene set/pathway analysis. These approaches integrate biological knowledge across omics levels through expert driven and computationally derived knowledge bases. This is a way to perform integrative analysis even when only a single omics level has been collected for a particular dataset. The knowledge bases incorporate and integrate data from a variety of omics levels to aid in systematic understanding. Pathway analysis methods can test whether the effects observed are enriched for various biological functions. These methods range in both complexity of statistical methods and the level of detail required to conduct the analysis. Relatively simple approaches, such as overrepresentation analyses (ORA), only require a set of statistically significant endpoints (e.g., genes, metabolites, and proteins) that test for enrichment in a set of endpoints known to be related to a biological process [10]. Slightly more complex approaches, such as GSEA [11] or the Correlated Lancaster Approach [5], use all of the data as either ranks or test statistics to determine if enrichment exists. These methods use all available data, addressing the limitation of ORA approaches which rely on an arbitrary significance threshold. Since many available databases contain more information than just groups of endpoints (e.g., genes, metabolites), incorporating information, such as pathway topology, will ultimately be desirable. Although the best way to incorporate these relationships is still an active research area, some methods (i.e., impact factor analysis) are currently able to leverage this information [10]. A full discussion of pathway and gene set analysis methods is beyond the scope of the current paper; an excellent review of the methods available is included in Ge et al. [14].

2.5. Replication as a Form of Integration. Methods development is an incredibly active area of research, and promising new methods for integrating “-omics” data are on the horizon. The simplest of these approaches use different “-omics” technologies as “pseudo replication” across “-omes,” building on the simple overlap approaches discussed above. Specifically, sophisticated Bayesian approaches incorporate information from one “ome” as prior information to perform association analysis for other, distinct “omes.” Additionally, there are a number of clustering and network based analysis tools that do not rely on established knowledge bases and have the potential to discover new biology. However, these approaches have not been widely used due to the limited number of datasets amenable to such analysis [12, 13]. An excellent review of newer approaches for “omics” integration is included in Ge et al. [14].

2.6. Constantly Evolving Methods. Although methods and data integration techniques are continuously evolving, there are several challenges that will need to be addressed in order for an integrative approach to become standardized and routine. From a statistical perspective, the most fundamental challenge in integrative analyses is dimensionality: taking more levels into account in the analysis tends to increase the dimensionality of the problem. Adding more layers of data or increasing the resolution of measurements increases the dimension of unknown parameters, which are often difficult to estimate, thereby making the overall inference weaker. This might seem paradoxical, as the purpose of taking multiple levels into account is precisely the opposite—to use more observations to obtain a more accurate picture of the biological system under study. In addition to the challenges described above with high dimensional data, a formidable quandary is how best to link data across omics platforms and different levels of biology. These relationships often produce “one-to-many” relationships, making causal relationships difficult to define.

There are also a number of limitations in the data curation and quality control. In addition, at every step, there will be checkpoints of compatibility of the data, such as normalization to the same scale, sample selection from representative cohorts, adequate correction for technical batch effects, and use of different platforms. Constantly evolving technologies exacerbate this challenge. Additionally, the variety of study designs underlying individual “-omics” datasets poses a significant problem to integrating data across multiple studies. Many studies are cross-sectional and only capture a snapshot of what is actually a highly dynamic system (e.g., transcriptomics). In addition, cohort differences due to individual study goals or available study populations may pose a significant barrier for integrating data. Even if the study designs are similar, different technology platforms have varying resolutions. Although genotype imputation has reduced this barrier in genome-wide association studies, a comparable tool does not exist for most “-omics” platforms. Pathway databases are continuously improving, but currently information related to tissue type, cell type, developmental stage (young, old), and disease state is extremely sparse. We need methodological improvements that can address pathway topologies and feedback loops and methods to simulate data to benchmark pathway analysis methods.

2.7. Comparative Approaches for Integrating Omics Data. Another aspect of data integration and “-omics” integration can be seen in the growing reliance on in vitro model systems and comparative genetics approaches with model organisms of disease. To date, there have been many successful examples of comparative genomics that implement a multiomics strategy to validate or replicate signals. For example, lymphoblastoid cell line models have shown success in finding gene and gene expression results that support clinical
genetic results [15]. Results across species, such as domestic dogs, have also been shown to be excellent approaches for omics integration [16, p. 1], [17]. Cancer research is an example where a comparative approach using canines holds particular promise [18]. Comparisons of DNA copy-number aberrations in canines and human have provided valuable insight into the mechanisms of osteosarcoma, lymphoma, intracranial tumors, and other cancer types [17, 19, 20].

Although it will take a tremendous effort from the research community to address many of these remaining challenges to “-omics” integration, there are useful methods currently available and no shortage of available “-omics” data. With the massive amounts of data that we have and are currently being generated, the challenge will now be to integrate these technologies to form a cohesive biological depiction of human disease, because only then can we claim to have considered all constituents collectively.

An enormous challenge is also the functional validation of the in silico findings in relevant living biological systems, as well as the development of adequate in vitro functional studies to keep up with the increasing throughput by which candidates for validation are generated. It is still crucial to explore functions of thousands of candidate genes, proteins, and metabolomics to ascertain their value as risk factors, as predictive factors for therapy response, and as therapeutic targets.

3. Integrative Omics in Personalized Medicine

While integrative analyses are important in all areas of genetics and genomics, they are especially important for mapping and obtaining a better understanding for the biology of drug response. There are a number of study design limitations in pharmacogenomics, pharmacoproteomics, and pharmacometabolomics that force the use of integrative approaches to make reliable discoveries. In most omics applications, replication is considered the gold standard—where potential associations are tested in one dataset and significant signals are validated in independent data. Obtaining and properly using these data are a particular challenge for drug response studies due to limitations in study design; for example, such studies are frequently nested within clinical trials, where sample sizes are extremely limited or treatment strategies are not completely comparable to the initial study. Because of this, replication samples are not routinely available and replication across omics levels is the only available option for reinforcing the initial discoveries.

Other opportunities for integrative approaches to address the needs in personalized medicine are in health monitoring. An example can be found in Chen et al. [21], where the authors develop an integrative personal omics profile (iPOP) analysis tool that tracks individual genomic, transcriptomic, proteomic, metabolomic, and autoantibody profiles [21]. This technology is successfully leveraged to identify healthy and diseased states for a single individual [21]. These approaches are in their infancy but provide great hope for the management and prevention of complex disease. Additional, integrative approaches specific for personalized medicine are thoroughly reviewed in Chen and Snyder [22].

4. Conclusions

A more fundamental understanding of the biological dynamics across omics datasets will enable us to better identify risk factors, refine disease diagnosis, predict therapeutic effects and prognosis, and identify new targets for therapy in personalized medicine. While the biological intuition of integrative “-omics” is clear, the real challenges are related to data integration, curation, and analysis. As we are moving towards an era in which the amount of data produced every year is increasing exponentially, methods to develop a deeper understanding of the biology of complex systems are crucial.

Competing Interests

The authors declare that they have no competing interests.

References


Clinical Application of a Modular Genomics Technique in Systemic Lupus Erythematosus: Progress towards Precision Medicine

Eric Zollars, 1 Sean M. Courtney, 2 Bethany J. Wolf, 3 Norm Allaire, 4 Ann Ranger, 4 Gary Hardiman, 2 and Michelle Petri 5

1 Division of Rheumatology, Medical University of South Carolina, Charleston, SC 29425, USA
2 Center for Genomic Medicine, Medical University of South Carolina, Charleston, SC 29425, USA
3 Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC 29425, USA
4 Biogen Idec, Cambridge, MA 02142, USA
5 Division of Rheumatology, Johns Hopkins School of Medicine, Baltimore, MD 21287, USA

Correspondence should be addressed to Eric Zollars; zollars@musc.edu

Received 13 May 2016; Accepted 20 July 2016

Academic Editor: Brian Wigdahl

Copyright © 2016 Eric Zollars et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Monitoring disease activity in a complex, heterogeneous disease such as lupus is difficult. Both over- and undertreatment lead to damage. Current standard of care serologies are unreliable. Better measures of disease activity are necessary as we move into the era of precision medicine. We show here the use of a data-driven, modular approach to genomic biomarker development within lupus—specifically lupus nephritis.

1. Introduction

Systemic lupus erythematosus (SLE) is the prototypical autoimmune disease. Immune tolerance breaks down leading to the immune system attacking normal tissues. Antibodies form that recognize self-antigens and lead to pathologic immune complex deposition. Dysfunction within both the innate and adaptive immune systems leads to increased cytokine production, especially type I interferons, B-cell overproduction of autoantibodies, and T-cell enhancement of these processes. This leads to the host of clinical abnormalities within SLE that includes rashes, oral ulcers, arthritis, inflammation around the heart and lungs, and cytopenias as well as severe renal and neurologic manifestations.

Treatment of the disease requires attention to all of these possible manifestations of disease activity. Any patient, at any time, can develop increased disease activity in any organ. While some manifestations are obvious (thrombocytopenia) others can be more subtle and complicated by other potential etiologies (is the rash or joint pain from lupus or something else?). There are no serologies which are universally useful for all manifestations across all patients. Anti-dsDNA antibodies associate with renal disease activity in a subset of patients but in only a fraction of these patients do changes in the levels predict changes in disease activity [1]. This is true for complement consumption as well. An inflammatory disease is expected to have elevated inflammatory markers (erythrocyte sedimentation rate, ESR, C-reactive protein, and CRP). In SLE, large changes in ESR are informative (with the caveat that infection also raises the ESR) but minor fluctuations are not [2]. Further, many lupus patients maintain a baseline elevated ESR regardless of clinical disease activity. CRP is less often elevated in SLE outside of infection [3]. There is no shortage of attempts at producing better biomarkers for lupus disease activity. Novel autoantibodies (anti-C1q), cell surface markers, and cytokines are all reported [4]. Only complement surface deposition has reached a level of reproducibility to allow inclusion in a commercially available product (AVISE, http://www.exagen.com/).

Exploration of gene expression in SLE began in 2002 with the cytokine specific arrays in Rus et al. [5]. They
showed increased expression of inflammatory cytokines and were able to separate lupus patients from healthy controls. A seminal work within this area was by Baechler et al. in 2003 showing the presence of an interferon signature [6]. Further work reproduced the interferon signal but also a curious neutrophil related signature [7]. However, these early works revealed the well-described limitations of microarray work—difficulties with technical and biological variability [8]. Technical variability includes batch effects as well as different probe sequences in different platforms. Biological variability leads to the lack of reproducibility between microarray studies. Furthermore, the interferon signature developed in Baechler was not shown to track disease activity [9].

The limitations in large-scale gene expression analysis led to the development of gene ontology [10] and functional enrichment methods such as gene set enrichment analysis (GSEA) [11]. These methods involve the annotating of transcripts with known roles in biologic processes and pathways as well as molecular structure and cellular components. These methods were primarily performed within oncologic processes.

Chaussabel and coworkers developed a novel, data-driven method using primarily inflammatory diseases [12]. In this method, clusters of genes that were observed in multiple disease processes were clustered into "modules." These modules were then labeled based on their primary function based on an automated literature search. This method was shown to discriminate between active and inactive disease within a pediatric lupus cohort. The method, initially developed on Affymetrix arrays was updated for use on Illumina arrays and the module list was greatly expanded. Importantly, the interferon module from 2008 was split into three modules, two of which were responsive to changes in disease activity. Further work by the group, recently published [13], expanded the work further, showing longitudinal variation as well as treatment response. They showed the prominence of a plasma cell signature within a subset of pediatric lupus patients that was a reliable marker of disease activity.

The work here evaluates the utility of the modular approach in an adult lupus population. Banchereau et al. mention that the pediatric population is a special population to study and this is true as patients who present with lupus in the pediatric years are more likely to have severe disease such as lupus nephritis or neurologic disease. Furthermore, we are specifically evaluating how a modular approach can discriminate between active and inactive lupus patients. This is likely a more difficult problem than discriminating between lupus patients and healthy controls. Also, it is unclear whether differences found between lupus and healthy controls can be directly applied to differences within lupus patients.

2. Methods

2.1. Study Population and Design. The study protocol for SPARE (Study of biological Pathways, disease Activity and Response markers in patients with systemic lupus Erythematosus) was approved by the Johns Hopkins University School of Medicine Institutional Review Board. SLE patients were enrolled from the Hopkins Lupus Cohort following informed consent. Adult patients were eligible if they were aged 18 to 75 years and met the definition of SLE as defined by the revised American College of Rheumatology classification criteria [14]. At entry into the study the patient's medical history was reviewed and information on current medications was recorded. Visits were scheduled quarterly or more often if required for disease activity over a 2-year period. All patients were evaluated by the same physician at entry and all subsequent cohort visits (MP). Three hundred and six SLE patients were enrolled in the observational study. Patients were treated according to standard clinical practice. To assess disease activity, the Safety of Estrogens in Lupus Erythematosus: National Assessment (SELENA) version of the Systemic Lupus Erythematosus Disease Activity Index (SLEDAI) [15] as well as the Physician Global Assessment [16] was completed at each visit.

2.2. Sample Selection. The samples selected for this analysis were chosen retrospectively based on the recorded clinical information. The goal was to compare lupus patients with active disease to lupus patients with clinically quiescent disease. Lupus nephritis is one of the more objective and persistent forms of lupus disease activity. The amount of protein in the urine is a relatively reliable indicator of ongoing inflammation within the kidney (with caveats, [17]). The "high activity" patients were selected based on the presence of a urine protein/creatinine ratio of 1.6 or higher. There was no selection for medications, ethnicity, or age. The "no activity" patients were clinically assessed to have no lupus disease activity (PGA = 0) and were on no immune-modifying medications other than hydroxychloroquine. Standard practice in lupus treatment includes continuing hydroxychloroquine therapy regardless of disease activity. Another comparator group, "typical" lupus, was created to compare specifically to the healthy controls. This group was randomly chosen from the dataset with the only restriction that no patient would be represented more than once. The healthy control patients were collected by Biogen and were never assessed by MP. Patient characteristics are shown in Table 1.

2.3. Sample Preparation. Peripheral blood samples used for gene expression analyses were collected using the PAXgene Blood RNA system (PreAnalytix GmbH). RNA was isolated from PAXgene preserved blood using the Agencourt RNAdvance Blood kit automated on an Arrayplex liquid handling system (Beckman Coulter, Indianapolis, IN). RNA integrity and concentration were assessed using the HT RNA reagent kit (Caliper Life Sciences, Hopkinton, MA) using a LabChip GX (PerkinElmer, Waltham, MA). RNA samples with a RQS score of >8.0 were considered of acceptable quality for downstream applications.

2.4. Gene Expression Analysis. RNA (50 ng) isolated from the PAXgene blood sample was amplified and biotin-labeled with the NuGEN Ovation RNA Amplification system V2, Ovation WB reagent, and Encore Biotin module (NuGEN Technologies, Inc., San Carlos, CA) using an Arrayplex automated liquid handler (Beckman Coulter, Indianapolis, IN). 2 ug of...
biotin labeled ssDNA probe was hybridized to Affymetrix GeneChip HT HG-U133+ PM plate arrays with modified conditions as described in Allaire et al. [18]. Washing and staining of the hybridized arrays were completed as described in the GeneChip Expression analysis technical manual for HT plate arrays using the Genechip® Array Station (Affymetrix, Santa Clara, CA) with modifications as described in Allaire et al. [18]. The processed Genechip plate arrays were scanned using GeneTitan scanner (Affymetrix, Santa Clara, CA). Affymetrix scans were subjected to standard quality control (QC) measures. These tests included a visual inspection of the distribution of raw signal intensities and an assessment of RNA degradation, relative log expression (RLE), and normalized unscaled standard error (NUSE). All sample scans passed these QC metrics. CEL files were subjected to GC-content-based Robust Multi-array Average (GCRMA) normalization [19]. Expression levels were log (base 2) transformed. All calculations and analyses were carried out using R and Bioconductor computational tools [20]. Modules used included limma [21], PAMR [22], and GEOQuery [23]. TopGene (https://topgene.cchmc.org/) was used for functional enrichment analysis. Genes composing Chaussabel modules were taken directly from Table S2 in [12].

3. Results

An analysis of differential gene expression leads to a total of 799 genes differentially expressed at a Benjamini-Hochberg adjusted significance of 0.05. There is an obvious difference between the high and no disease activity groups as seen in the heatmap in Figure 1. For the most part, the high disease activity clusters to the left and the low disease activity clusters to the right. Simple hierarchical clustering separates 10/13 high disease activity from the no disease activity.

Functional enrichment analysis using TopFunn shows results that would be expected for lupus (Table 2). There is significant upregulation of type I interferon pathways, immune pathways, and cytokine-associated pathways. There was nothing unexpected in this analysis.

A search for a “gene signature” that would separate high lupus disease activity from no disease activity used PAMR. This is a clustering method that finds the smallest list of genes that leads to the smallest misclassification error. These genes are shown in Table 2. While there are genes in the list that are biologically plausible, it suffers from “noise” inherent in these microarray gene lists [24].

We next looked at the modules developed by Chaussabel and colleagues. We used the 2008 modules as they were developed using the Affymetrix platform. The arrays used in the study reported here lack the mismatch probes of the U133A and U133B chips used by Chaussabel but are otherwise the same probes. We first reproduced the work of Chaussabel using the datasets available on the NIH GEObus (GSE11909). These were comparing untreated pediatric lupus to healthy controls. This dataset was not complete, missing 6 of 12 U133B chips of the healthy controls, but all lupus and all 133A chips were available. This is shown in Table 3. Each module has a certain number of genes. For instance, Module 1.1, plasma cells, is defined by the membership of 76 genes. The first data box in Table 3 is reproduction of the pediatric lupus data.
Table 2: Functional enrichment analysis from ToppGene. Hit count in query list is the number of genes from the list of significantly differentially expressed genes in that ontology. Hit count in genome is the number of described genes in the ontology.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>p value</th>
<th>q-value</th>
<th>Hit count in query list</th>
<th>Hit count in genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006955</td>
<td>Immune response</td>
<td>$1.35E-25$</td>
<td>$6.46E-22$</td>
<td>114</td>
<td>1416</td>
</tr>
<tr>
<td>GO:0019058</td>
<td>Viral life cycle</td>
<td>$9.43E-24$</td>
<td>$4.50E-20$</td>
<td>50</td>
<td>314</td>
</tr>
<tr>
<td>GO:0045087</td>
<td>Innate immune response</td>
<td>$2.20E-23$</td>
<td>$1.05E-19$</td>
<td>84</td>
<td>883</td>
</tr>
<tr>
<td>GO:0044764</td>
<td>Multiorganism cellular process</td>
<td>$4.44E-23$</td>
<td>$2.12E-19$</td>
<td>75</td>
<td>725</td>
</tr>
<tr>
<td>GO:0071357</td>
<td>Cellular response to type 1 interferon</td>
<td>$4.70E-23$</td>
<td>$2.25E-19$</td>
<td>27</td>
<td>76</td>
</tr>
<tr>
<td>GO:0060337</td>
<td>Type I interferon signaling pathway</td>
<td>$4.70E-23$</td>
<td>$2.25E-19$</td>
<td>27</td>
<td>76</td>
</tr>
<tr>
<td>GO:0051607</td>
<td>Defense response to virus</td>
<td>$6.08E-23$</td>
<td>$2.90E-19$</td>
<td>41</td>
<td>212</td>
</tr>
<tr>
<td>GO:0034340</td>
<td>Response to type 1 interferon</td>
<td>$7.06E-23$</td>
<td>$3.37E-19$</td>
<td>27</td>
<td>77</td>
</tr>
<tr>
<td>GO:0016032</td>
<td>Viral process</td>
<td>$3.62E-22$</td>
<td>$1.73E-18$</td>
<td>73</td>
<td>714</td>
</tr>
<tr>
<td>GO:0009615</td>
<td>Response to virus</td>
<td>$1.85E-21$</td>
<td>$8.84E-18$</td>
<td>47</td>
<td>310</td>
</tr>
<tr>
<td>GO:0002252</td>
<td>Immune effector process</td>
<td>$2.18E-21$</td>
<td>$1.04E-17$</td>
<td>67</td>
<td>628</td>
</tr>
<tr>
<td>GO:0006952</td>
<td>Defense response to other organisms</td>
<td>$1.45E-19$</td>
<td>$6.93E-16$</td>
<td>107</td>
<td>1515</td>
</tr>
<tr>
<td>GO:0098542</td>
<td>Defense response to other organisms</td>
<td>$5.35E-19$</td>
<td>$2.56E-15$</td>
<td>50</td>
<td>401</td>
</tr>
<tr>
<td>GO:0043207</td>
<td>Response to external biotic stimulus</td>
<td>$1.29E-18$</td>
<td>$6.14E-15$</td>
<td>68</td>
<td>726</td>
</tr>
<tr>
<td>GO:0031707</td>
<td>Response to other organisms</td>
<td>$1.29E-18$</td>
<td>$6.14E-15$</td>
<td>68</td>
<td>726</td>
</tr>
<tr>
<td>GO:0009607</td>
<td>Response to biotic stimulus</td>
<td>$1.43E-17$</td>
<td>$6.82E-14$</td>
<td>68</td>
<td>760</td>
</tr>
<tr>
<td>GO:0048525</td>
<td>Negative regulation of viral process</td>
<td>$1.43E-16$</td>
<td>$6.85E-13$</td>
<td>22</td>
<td>78</td>
</tr>
<tr>
<td>GO:0034097</td>
<td>Response to cytokine</td>
<td>$2.85E-16$</td>
<td>$1.36E-12$</td>
<td>59</td>
<td>629</td>
</tr>
<tr>
<td>GO:0071345</td>
<td>Cellular response to cytokine stimulus</td>
<td>$5.70E-16$</td>
<td>$2.72E-12$</td>
<td>53</td>
<td>527</td>
</tr>
<tr>
<td>GO:0006414</td>
<td>Translational elongation</td>
<td>$2.32E-15$</td>
<td>$1.11E-11$</td>
<td>26</td>
<td>130</td>
</tr>
<tr>
<td>GO:0045069</td>
<td>Regulation of viral genome replication</td>
<td>$1.31E-14$</td>
<td>$6.27E-11$</td>
<td>18</td>
<td>58</td>
</tr>
<tr>
<td>GO:0006413</td>
<td>Translational initiation</td>
<td>$1.87E-14$</td>
<td>$8.94E-11$</td>
<td>29</td>
<td>179</td>
</tr>
<tr>
<td>GO:0019221</td>
<td>Cytokine-mediated signaling pathway</td>
<td>$4.42E-14$</td>
<td>$2.11E-10$</td>
<td>43</td>
<td>402</td>
</tr>
<tr>
<td>GO:0039079</td>
<td>Viral genome replication</td>
<td>$1.94E-13$</td>
<td>$9.25E-10$</td>
<td>19</td>
<td>76</td>
</tr>
<tr>
<td>GO:0043900</td>
<td>Regulation of multiorganism process</td>
<td>$4.19E-13$</td>
<td>$2.00E-09$</td>
<td>37</td>
<td>325</td>
</tr>
<tr>
<td>GO:0035455</td>
<td>Response to interferon-alpha</td>
<td>$4.56E-13$</td>
<td>$2.18E-09$</td>
<td>11</td>
<td>19</td>
</tr>
</tbody>
</table>

The number of significantly expressed genes is shown as a proportion. Thus, 45% of the genes within Module 1.1 are increased in expression when comparing untreated pediatric lupus to healthy controls. This nearly exactly reproduces the original work as expected. This work showed the notable increased expression of genes within the interferon, plasma cell, neutrophil, erythrocyte, and myeloid modules. Decreased expression was seen in genes associated with the ribosomal, cytotoxic, and T-cell modules.

We then looked at the modular representation of gene expression differences between high and low disease activity in adult SLE patients. The reason for doing this is that we hypothesized that what distinguishes high and low disease lupus within SLE may be different than what distinguishes SLE from healthy controls. In the second data box, “Adult SLE, High versus No Activity,” we show that 97% of genes within the interferon module are significantly increased in expression. Thus, even within SLE there is increased expression of the genes that make up the interferon module. This is similar to the untreated pediatric lupus population. It is also quite expected as many studies have shown the importance of the interferon pathways in SLE. Remarkably, even in this treated, adult population we see an increase in the neutrophil signature. The neutrophil signature is associated with lupus nephritis [25] and this was also demonstrated in the most recent pediatric SLE study [13]. We do not observe
Table 3: Module name and number as described in [12]. The “pediatric SLE” column is reproduced from GSE11909. The numbers shown are the percentage of genes from the module that have significant differential gene expression (DGE). Thus, the first row is interpreted as follows: 45% of the genes in Module 1.1 have significant DGE and are increased in expression. The second column, “Adult SLE, High versus No Activity,” is a DGE analysis between lupus patients with high disease activity and lupus patients with no activity. The third column, “Adult SLE versus HC” is a DGE analysis between a cohort of lupus patients with average disease activity and healthy controls.

<table>
<thead>
<tr>
<th>Module name</th>
<th>Module number</th>
<th>Pediatric SLE</th>
<th>Adult SLE</th>
<th>Adult SLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Untreated versus HC</td>
<td>High versus no activity</td>
<td>SLE versus HC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Up</td>
<td>Down</td>
<td>Up</td>
</tr>
<tr>
<td>Plasma cells</td>
<td>1.1</td>
<td>45</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Erythrocytes</td>
<td>1.2</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>B-cells</td>
<td>1.3</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>None</td>
<td>1.4</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Myeloid</td>
<td>1.5</td>
<td>13</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>None</td>
<td>1.6</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Ribosomal</td>
<td>1.7</td>
<td>0</td>
<td>78</td>
<td>5</td>
</tr>
<tr>
<td>None</td>
<td>1.8</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Cytotoxic</td>
<td>2.1</td>
<td>2</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>Neutrophils</td>
<td>2.2</td>
<td>33</td>
<td>0</td>
<td>55</td>
</tr>
<tr>
<td>Erythrocytes</td>
<td>2.3</td>
<td>26</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Ribosomal</td>
<td>2.4</td>
<td>0</td>
<td>77</td>
<td>0</td>
</tr>
<tr>
<td>None</td>
<td>2.5</td>
<td>1</td>
<td>27</td>
<td>3</td>
</tr>
<tr>
<td>Myeloid</td>
<td>2.6</td>
<td>23</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>None</td>
<td>2.7</td>
<td>1</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>T-cells</td>
<td>2.8</td>
<td>0</td>
<td>42</td>
<td>0</td>
</tr>
<tr>
<td>None</td>
<td>2.9</td>
<td>3</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>None</td>
<td>2.1</td>
<td>4</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>None</td>
<td>2.11</td>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Interferon</td>
<td>3.1</td>
<td>91</td>
<td>0</td>
<td>97</td>
</tr>
<tr>
<td>Inflammation</td>
<td>3.2</td>
<td>13</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>Inflammation</td>
<td>3.3</td>
<td>9</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>None</td>
<td>3.4</td>
<td>6</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>None</td>
<td>3.5</td>
<td>14</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>None</td>
<td>3.6</td>
<td>6</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>None</td>
<td>3.7</td>
<td>4</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>None</td>
<td>3.8</td>
<td>1</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>None</td>
<td>3.9</td>
<td>2</td>
<td>11</td>
<td>1</td>
</tr>
</tbody>
</table>

significantly increased numbers of genes associated with the plasmablast signature.

Next we looked at a cohort of adult lupus patients with a typical mixture of disease activity and compared to healthy controls. This is an attempt to make the comparison with Data Box 1 but with adult lupus patients on a mix of therapies. Characteristics of this group are shown in Table 1. The average SLEDAI for this group is 2.7, representing mild disease. Only eight of the 95 patients have renal disease, six have arthritis, and none have significant neurologic disease. Again, we do not see the plasma cell signature reported in Chaussabel et al. [12] and Banchereau et al. [13]. Again, the interferon signal is very strong but notably is not significantly different from either of the other comparisons. There are notable differences between the comparison within lupus patients and the comparison between lupus and healthy controls. For the neutrophil module, 55% of the genes were present in high disease activity while only 29% showed this difference when comparing “typical lupus” to healthy controls. This is roughly what was seen in the pediatric population where there was some evidence that increased expression of genes in the neutrophil module associated with increased disease activity. Increased presence of reduced expression genes was seen in the two ribosomal modules, T-cell module and cytotoxic modules.

4. Discussion

The clinical evaluation and treatment of patients with SLE is in desperate need of advanced biomarker development. Assessment of disease activity is difficult and currently inadequate. Medical treatment decisions are, for the most
part, not guided by individual characteristics of the patient or the disease. If we are ever to achieve precision medicine in this complex, heterogeneous disease it will be through detailed molecular phenotyping and close monitoring of reliable indicators of disease activity.

Gene expression analysis allows for measurement of many variables at once, potentially allowing for capture of the heterogeneity of this complex disease. Multiple techniques for dimension reduction have been proposed and one of the more promising for the quantification of disease activity in lupus is the modules developed in pediatric lupus by Chaussabel et al. [12]. We show here the application of these modules in quantification of SLE disease activity, specifically lupus nephritis.

The results of this analysis show some similarities with the pediatric lupus patients studied in Dallas [12, 13, 26]. First, the prominence of the interferon signature is reproduced. This is not surprising based on the fundamental importance of interferon in SLE pathology. It is worth noting that a “score” of 97% in the interferon module indicates that 97% of genes in that module were found to have significant differential gene expression. Comparing SLE patients with high disease activity to those with low disease activity reveals increased expression of these genes. However, comparing “typical” SLE activity to healthy controls reveals increased expression as well. The score does not indicate the level of gene expression only that there was an increase from one group to another. Microarray measurement of gene expression has some association with more quantitative methods of measuring mRNA, for example qPCR, but is not as sensitive to changes in expression [27]. We did not make an attempt here to further quantify the amount of expression within the module. Further work in this area can include an analysis with RNaseq that leads to actual mRNA counts.

Another similarity with the pediatric study is the presence of the neutrophil module. Truly remarkable work evaluating the roles of neutrophils in SLE pathology emerged after the demonstration of the neutrophil signature [7]. Unlike the interferon module above, the neutrophil module did show increased numbers of differentially expressed genes in the active-inactive group compared to the lupus-healthy control group. This is possibly due to the demonstrated role of neutrophils in lupus nephritis specifically [25]. However, it is also possible that multiple neutrophil-related interactions are involved. Perhaps increased disease leads to recruitment of more of these pathways or that some patients have different components of these pathways.

A significant difference between this work and the original work of Chaussabel is in the plasma cell module. Plasma cells are the antibody producing cells and are increased in active lupus. In the recent work with the Dallas pediatric SLE cohort the plasmablast module was found to associate with disease activity and was significantly reduced by treatment with mycophenolate [13]. We find here, in an adult population, the absence of a significant plasma module differentiating either active from inactive disease or typical SLE from healthy controls. This seems to be also true in the primarily Caucasian, French population analyzed in Chiche et al. [26]. In that work, the newer Illumina-based modules were used where the plasma cell module is Module 4.11. Thus, it is unclear at this point if the plasma cell signature is enhanced in the pediatric population studied or unable to be differentiated from treatment effects in this study.

This study has multiple limitations. First, for a true biomarker study the response to change in disease activity and outcomes would have to be followed over time. At this time we were interested in the applicability of the modules in a specific subpopulation of lupus but plan for further analysis. The limited size of this sample did not allow for treatment effects to be studied, though great effort was made to minimize those differences in the active versus inactive group. Finally, one of the limitations is that the technology of the Affymetrix microarray is limiting and future work will include more quantitative and reproducible techniques.

There is great promise in the use of data-driven analysis in the exploration of complex, heterogeneous diseases such as lupus. We show here an example of how this can be used in evaluating active versus inactive disease within SLE. As we move to precision medicine methods such as these will lead to better characterization of disease, better therapies, and better response to therapies.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The Hopkins Lupus Cohort is supported by NIH AR43727. Research reported in this publication was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under Award no. T32AR048522 as well as the Ira T. Fine Discovery Fund (to Eric Zollars).

References

[6] E. C. Baechler, F. M. Batiwalla, G. Karypis et al., “Interferon-inducible gene expression signature in peripheral blood cells of...


