Classification of Tissues and Disease Subtypes using Whole-Genome Signatures

A Thesis

Submitted to the Faculty

of

Drexel University

by

Michael P. Gormley

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

August 2008

Table of Contents

List of Tables

List of Figures

Abstract

Classification of Tissues and Disease Subtypes Using Whole-Genome Signatures

Michael P. Gormley
Aydin Tozeren, Ph. D.

Development and application of microarray technology in biological research has led to compilation of expression and sequence data on a genome-wide scale. Given the volume of data produced and the complexity of gene regulatory mechanisms, it can be difficult to extract meaningful biological information. Classification can be used to reduce the complexity through the detection of genes, genetic loci or conditions that share common attributes and the identification of gene expression patterns or genotypes associated with phenotype. In the study of cancer, supervised classification has been applied to identify gene expression biomarkers of different disease states. Clinically validated biomarkers are valuable indicators for diagnosis and guiding therapeutic strategy. We developed an iterative machine learning algorithm to compare the predictive value of biomarker sets chosen by supervised classification against sets selected randomly from known disease-related genes. Both supervised classification and feature selection based on prior knowledge resulted in discriminative classification of molecular phenotypes in breast cancer and lymphoma. Compilation of gene expression data has led to the identification of genes with bimodal, or switch-like, expression patterns. We used unsupervised, supervised and model-based classification methods to investigate the biological relevance of bimodal expression patterns and to evaluate their potential for class discovery and prediction. Both model-based and supervised classification resulted in the accurate classification of samples by tissue phenotype or infectious disease. Functional enrichment analysis indicates switch-like genes are involved in tissue-specific or immune response functions. Taken together, this evidence supports the assertion that bimodal expression patterns are biologically relevant. Clinical relevance of bimodal expression patterns was investigated in an association study of genotypes of families affected by autism. A subset of neural-specific switch-like genes was used to identify candidate gene regions which may contain genetic variants associated with autism risk. A two-stage family-based association test detected an autism susceptibility locus in the q26 region of chromosome 10. The coding region of the fibroblast growth factor receptor 2 (FGFR2) gene is 80 kilobases downstream from the identified locus. Altered expression of FGFR2 may be a contributing genetic factor in development of autism. Identification of the susceptibility locus provides motivation for novel hypotheses concerning the molecular basis of autism. In addition, we provide a method for integration of gene expression and genotype data that may lead to the identification of disease-related polymorphisms in other disorders.

Chapter 1: Introduction

Gene expression is regulated by complex interactions between DNA, regulatory proteins, epigenetic mechanisms and microRNA molecules. Activation and repression of gene expression is used to control cellular processes and can also lead to phenotypic changes such as tissue differentiation and disease. Microarray technology provides the means to quantify expression and type genetic variation in the DNA sequence at a genome-wide scale. However, the high-dimensionality of microarray datasets along with relatively small samples sizes hinders the effectiveness of microarray analysis. In this work, we have used a variety of classification methodologies to address critical issues in the field of microarray analysis and extract meaningful biological information.

Gene expression biomarkers are highly valued in the prediction of prognosis of heterogeneous disease. Supervised classification methods, such as k-nearest neighbor, linear discriminant analysis and support-vector machines, have been applied to gene expression microarray data in order to identify biomarkers at high-throughput. The lack of commonly shared genes among independent biomarker panels of the same disease state raises questions concerning the power and reproducibility of differential expression analysis. To address these questions, we developed a machine learning algorithm to generate and validate populations of gene expression biomarker panels from microarray data. With this approach, we identified many gene sets that are predictive of molecular subtype in breast cancer and lymphoma. In addition, we observed that the accuracy of classification decreases and the variance in accuracy increases when evaluating biomarker sets across platforms. From this analysis, we conclude that the lack of agreement between independently derived biomarker panels is due in part to the number of relevant genes and technical variation between microarray platforms. These results have important consequences in the experimental design and interpretation of microarray experiments.

Gene expression profiling of diverse phenotypes in health and disease allow us to identify common modes of gene expression. For example, housekeeping genes have been identified which are constitutively expressed across tissues and tend to be involved in a minimal set of structures and processes required for cellular viability. In previous work, we identified a set of switch-like genes with bimodal expression patterns across 19 different tissue types. In this work, we investigated the expression profiles of these switch-like genes in both health and infectious disease. Both model-based and multi-class supervised classification accurately categorized tissue samples according to tissue type and infectious disease. In addition, functional enrichment analysis indicated that activated switch-like genes in different phenotypes are involved in specialized tissue-specific or immune response functions. Through our application of advanced classification algorithms along with the use of gene functional information, we conclude that switch-like genes represent a biologically relevant subset of genes that warrant further study.   In addition, due to the accurate classification of phenotypes in a multi-class setting, we contend that the identification of switch-like genes may be a useful dimension reduction method in future analysis of expression data.

Use of gene sequencing arrays for the identification of single nucleotide polymorphisms associated with susceptibility to disease is burdened by a large multiple testing problem. Thresholds for genome-wide significance must be adjusted to account for the number of hypotheses tested. We used our insight on switch-like genes to reduce the number of hypotheses up front by identifying candidate gene regions in an association study of autism. Specifically, we scanned the coding and *cis*-regulatory regions of neural-specific switch-like genes for genetic variants associated with autism susceptibility. Using a two-stage family-based association test, we identified an autism susceptibility locus in an intergenic region of chromosome 10. The locus is approximately 80 kilobases upstream of the fibroblast growth factor receptor 2 (FGFR2) gene. Fibroblast growth factor signaling is involved with both neurodevelopment and neural proliferation in the adult brain. Our results suggest that altered expression patterns of FGFR2 due to genetic variation at the autism susceptibility locus may contribute to increased risk of disease. This study provides a novel method for the integration of gene expression and gene sequence data in genome-wide association study. In addition, the identification of the autism susceptibility locus and the potential involvement of FGFR2 provides motivation for biologists to test new hypotheses regarding the molecular basis of autism.

Chapter 2: Background

This chapter describes biological processes related to gene regulation in health and disease. Microarray platforms capable of measuring gene expression and characterizing gene sequence are discussed. In addition, databases that have compiled information on genes and gene products are reviewed.

2.1 Molecular biology of gene regulation

Cellular function is governed by the production, interaction, modification and degradation of proteins generated from genetic information stored in the nucleus. Genetic information is stored as DNA, a highly ordered configuration of polymer strings of nucleotides arranged in a double helix formation. Association of DNA with histone protein complexes allows for tight packing and organization of the genetic information [1]. Genes and regulatory regions are identified by specific sequences of nucleotide bases. Alterations in the genetic sequence caused by copying errors, environmental effects [2], or viral infection [3] can result in the alteration of the structure and function of gene products. The central dogma of molecular biology describes the process by which genetic information encoded in the DNA sequence is first transcribed into RNA and then translated into polypeptides. This process is strictly regulated to control which gene products are expressed under which conditions.

Transcription is the production of RNA molecules from the coding regions of the DNA sequence. Transcription is initiated by the binding of a protein complex of basal transcription factors and RNA polymerase to the DNA at a specific recognition site upstream of the gene coding region [4]. Binding of additional transcription factors to upstream or downstream regulatory regions of the DNA amplifies the rate of transcription. Transcription factor activity can be regulated by several processes including protein synthesis, subcellular localization [5], ligand binding [6], dimerization [7], and phosphorylation [8]. In addition, the relationship of transcription factors to the genes they regulate is often many to many. These characteristics of transcriptional regulation allow for fine control of gene expression. Following transcription, synthesized RNA is processed into messenger RNA (mRNA) by removing non-coding regions, and capping the 5' and 3' ends of the transcript. Alternative splicing at this stage results in greater variation in the protein population and allows for additional regulation of gene expression [9].

Transcribed mRNA carries genetic information from the nucleus to the cytoplasm where it serves as a template in the formation of amino acid peptide chains that constitute the primary structure of proteins. Amino acids are represented by sets of nucleotide triplets in the mRNA sequence known as codons [4]. Additionally, start and stop codons signal for the initiation and termination of protein synthesis. Transfer RNAs are oligonucelotide molecules with sites for codon recognition and amino acid binding. Ribosomes induce translation by providing a site for the interaction of the mRNA transcript with transfer RNAs and catalyzing the formation of peptide bonds between sequential amino acids. Following translation, synthesized polypeptides fold into native, three-dimensional structures that confer protein activity.

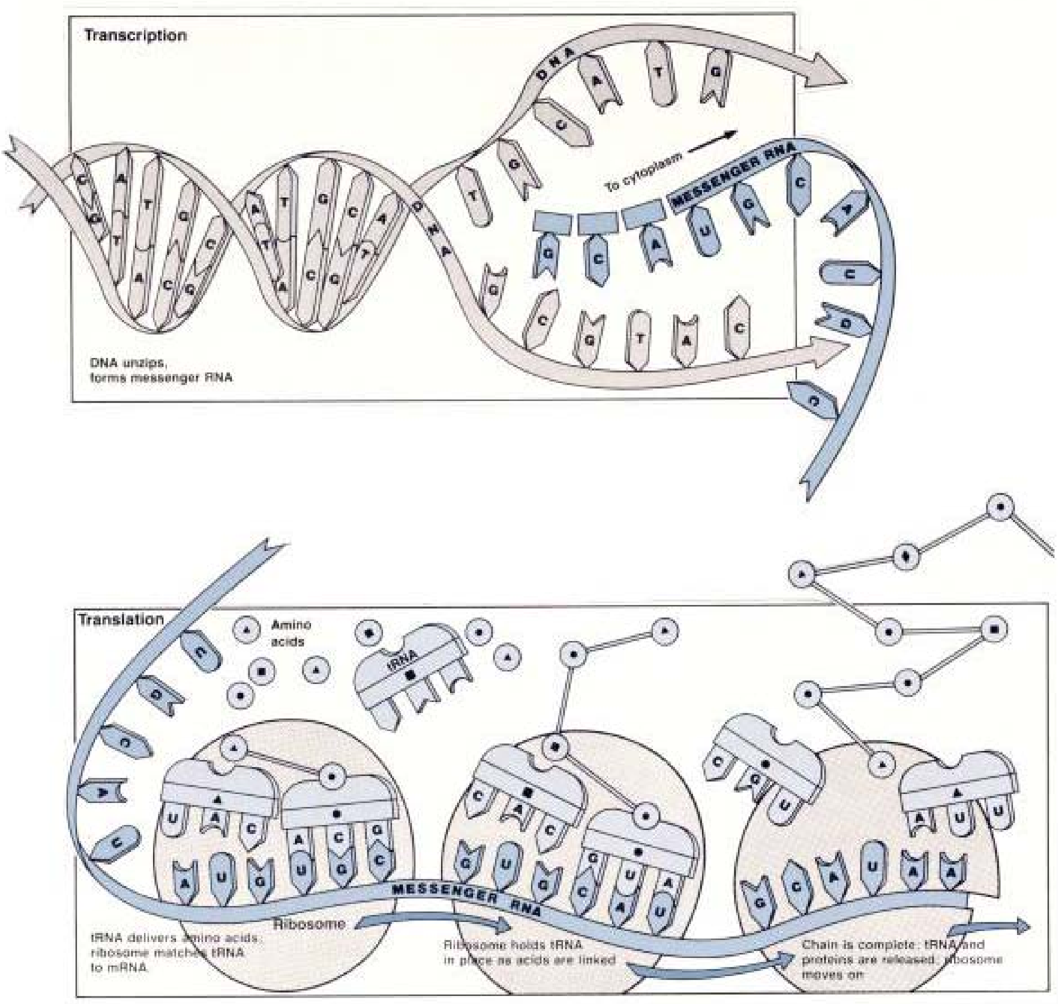


Figure 1: Transcription and translation [10]

Protein function can be altered by post-translational modifications. Modifications include the addition and removal of functional groups (e.g. phosphorylation, acetylation), covalent linkage to other proteins (e.g. ubiquitinylation) and the formation and cleavage of additional bonds between amino acids [11]. Kinases and phosphatases are enzymes which add or remove phosphate groups to serine, threonine or tyrosine residues of proteins [4]. Phosphorylation can change the activity of a protein through the blocking or formation of active sites or other conformational changes. Protein degradation is regulated by covalent binding of ubiquitin to lysine residues of the target protein. Ubiquitinylation marks proteins for degradation by proteolytic protein complexes. A number of proteins are synthesized in inactive forms. Activation occurs by formation or cleavage of bonds between amino acid residues [12]. Modifications such as these enable dynamic regulation of protein function at the post-translational stage.

Gene expression is also controlled by epigenetic mechanisms independent of gene sequence. Methylation of cytosine CpG dinucleotide sequences maintains gene expression patterns across cell division cycles and plays an important role in development [13, 14]. Distribution of CpG dinucelotides in the genome is disproportionately biased towards gene coding regions and transcription start sites [15]. DNA methylation surpresses transcription of associated genes through direct or indirect inhibition of transcription factor DNA binding [14]. Additional epigenetic mechanisms include histone modifications and chromatin remodeling. Phosphorylation, acetylation, and methylation of residues of the N-terminal histone tail alter the configuration of the DNA sequence to make it accessible for transcription [16]. Similarly, chromatin remodeling proteins use energy gained from ATP hydrolysis to dissociate DNA from histone complexes [16, 17]. Epigenetic mechanisms provide additional means of regulating the process of gene expression.

Association of mRNA transcripts with microRNAs (miRNA) provides regulation of gene expression at the post-transcriptional level. MicroRNAs are 21-25 nucelotide RNA molecules that bind with complimentary sequences in the 3’ untranslated region of mRNA transcripts [18]. Precursor microRNAs are transcribed in an inactive hairpin form. Production of mature microRNAs is catalyzed by two ribonuclease enzymes that yield the 21-25 base pair active form [19, 20]. Active microRNAs form a RNA-induced silencing complex (RISC) by association with proteins in the argonaute family [21]. In most conditions, binding of the RISC to the target mRNA represses gene expression by one of two mechanisms: either translational inhibition or destabilization of the target transcript through removal of the 3' polyamine cap [18]. Recent work has demonstrated that some miRNA-protein complexes may upregulate translation in growth arrested conditions [22]. Regardless of the effect on translation, miRNA-mediated effects result in transcript-specific regulation of expression.

2.2 Single-nucleotide polymorphisms

Single-nucleotide polymorphisms (SNPs) are variations in the genetic sequence consisting of a single nucleotide base substitution. Approximately four million SNPs distributed at an average density of one SNP per kilobase throughout the genome have been identified [23]. Genotyping efforts suggest that a majority of the genetic diversity between two individuals is captured by SNPs. Correlation between genetic variants arises as a result of common genetic history. Mutations are initially inherited together with alleles on the same chromosome. Linkage between alleles degrades over time by recombination and mutation, but extensive genotyping of single nucleotide polymorphisms in four genetically diverse populations through the International HapMap Project has identified common haplotypes (i.e. chromosomal loci that tend to be transmitted together) [HapMap, 2005; HapMap, 2007] [23, 24]. These studies have also located recombination hotspots where genetic variation occurs with higher frequency. This information can be used to genotype a large portion of the genetic sequence with a small number of tagSNPs that are highly associated with other SNPs in close proximity.

2.3 Cellular signaling pathways

Cells coordinate internal processes and respond to external stimuli through the use of cellular signaling pathways. Cellular signaling consists of cascades of biochemical reactions that act on proteins and small molecules to propagate messages through the cell to the nucleus. Many signaling pathways end in the activation of transcription factors or other DNA-binding proteins that then interact with the DNA to induce or repress expression of target genes.

The mitogen-activated protein kinase (MAPK) pathway is one of the most actively studied signaling pathways. MAPKs are divided into several major subgroups including extracellular signal-regulated kinases (ERK1, ERK2), c-Jun N-terminal kinases (JNK1, JNK2, JNK3) and stress-activated protein kinase-2 homologs (p38α, p38β, p38δ) [25, 26]. Signaling through ERK1/ERK2 pathways regulates cellular proliferation and cell division [26]. Activation of JNK pathways is involved with the initiation of apoptosis. The p38 MAPK family regulates cell division and gene expression related to osmotic and heat shock responses. In canonical MAPK signaling, MAPKs are activated by a phosphorylation cascade through a three-member protein kinase module [27]. For example, in ERK signaling, stimuli such as g-protein receptor (Ras) ligand binding induce the phosphorylation of an upstream MAP3K (a-Raf, b-Raf, c-Raf1)[28].

Activated MAPK3s phosphorylate MAP2Ks (MEK1/MEK2) at two serine residues. ERK1,2 is then activated through the phosphoryaltion of tyrosine and threonine residues. Upon activation, ERK1 translocates to the nucleus and activates transcription factors including c-Fos, ATF-2, Elk-1, c-Jun, c-Myc, and Ets1.

In addition to signal transduction along canonical pathways, a fair amount of crosstalk occurs between MAPK pathways and between MAPK and other canonical pathways. Activation of the p38 MAPK pathway inhibits activation of ERK signaling through the dephosphorylation of MEK1/2 [28]. In contrast, activation of integrin signaling through p21 protein-activated kinase 1 (PAK1) leads to the formation of focal adhesions and the phosphorylation of MEK1 [29]. PAK1-mediated phosphorylation enhances the association of MEK1 with Raf1 and leads to more efficient activation of ERK signaling [29]. Interactions along and between pathways form an elaborate signaling network by which cellular processes are integrated and controlled.

2.4 Mendelian and chronic diseases

Deregulation of cellular processes can lead to disease. Cancer, for example, is characterized by uncontrolled cell growth. Increasing chromosomal instability contributes to the accumulation of mutations that lead to advanced stages of disease. Mutation of the Ras proto-oncogene is observed in many tumors and causes constitutive activation of ERK signaling [26, 30]. Conversely, hypermethylation of CpG islands near gene promoters can cause silencing of tumor suppressor genes involved in DNA repair, hormone response, p53 signaling, apoptosis and cellular adhesion [31]. These findings

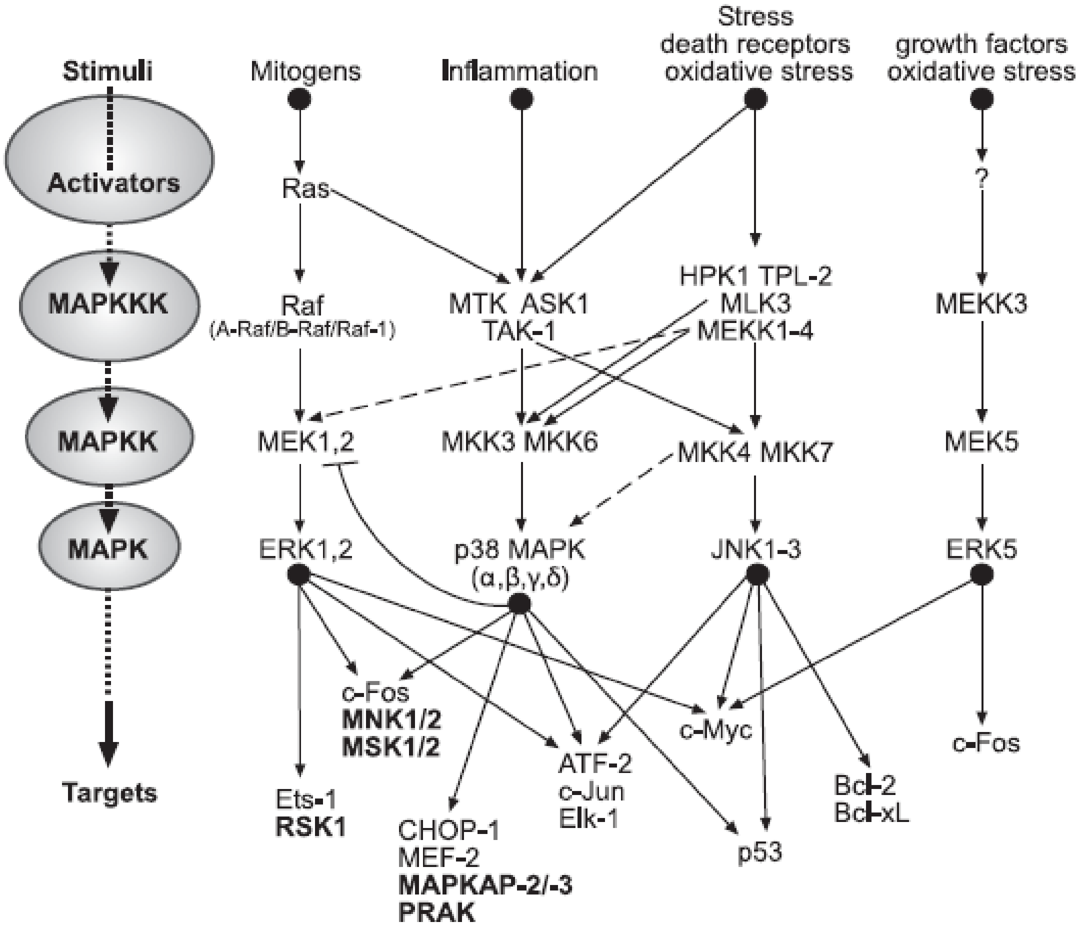


Figure 2: Crosstalk between mitogen-activated protein kinase (MAPK) signaling pathways. MAPK pathways are arranged into three-unit modules of protein kinases. Stimuli induce activation of downstream MAPKs through a cascade of phophorylation reactions. Context-specific crosstalk is indicated by dashed lines. [28]

demonstrate that disease is often associated with malfunctions in the cellular regulatory mechanism.

Diseases can be loosely grouped into two categories: mendelian and complex disorders. Mendelian, or monogenic, disorders are caused by the transmission of a defect at a single genetic locus [32, 33]. Examples of mendelian disorders include phenylketonuria, cystic fibrosis, Huntington disease, a subset of muscular dystrophies and genes that transmit heritable susceptibility to breast cancer and retinoblastoma [33]. Approximately 1200 mendelian disease genes have been identified through statistical analysis of the genotypes of families that demonstrate the disease phenotype [33]. Once a disease gene has been identified, hypotheses regarding the disease mechanism can be generated based on the function of the corresponding protein. Mendelian diseases are relatively rare. Complex disorders, such as cancer, obesity, autism and diabetes, are much more common. The common variant-common disease hypothesis proposes that susceptibility to diseases such as heart disease and cancer may be conferred in part by SNPs observed in the majority of the population. In support of this hypothesis, genome-wide association studies have identified disease-associated SNPs in many common diseases including rheumatoid arthritis, bipolar disorder, coronary artery disease, Crohn’s disease, hypertension, and type 1 and type 2 diabetes [34]. Complex disorders are associated with variation at many genetic loci as well as environmental and lifestyle factors. The contribution of multiple factors to the onset and progression of complex disorders makes it more difficult to isolate the genetic component of etiology.

2.5 Microarrays for gene expression and genotyping

Development of microarray platforms for genotyping and measuring gene expression has provided a powerful tool in the study of complex disease and other biological phenomena. Arrays consist of ordered arrangements of single-stranded oligonucleotides (probes) bound to a solid substrate such as a glass slide or plastic chip. Current technology provides for the representation of over 1 million features on a single array.

Microarray analysis enables genotyping and the quantification of gene expression on a near to genome-wide scale.

Gene expression microarrays can be classified into two categories: two-color complimentary DNA (cDNA) platforms and high-density oligonucleotide arrays. Two-color arrays are generally made in-house by spotting PCR-amplified cDNA probes with a robotic arrayer [35]. These arrays are used to quantify the relative gene expression in a set of samples. Briefly, mRNA transcripts (target) from a pair of samples are reverse-transcribed into cDNA and labeled with red and green fluorescent dyes.

Labeled sample cDNA is hybridized to the array and the relative abundance of mRNA in

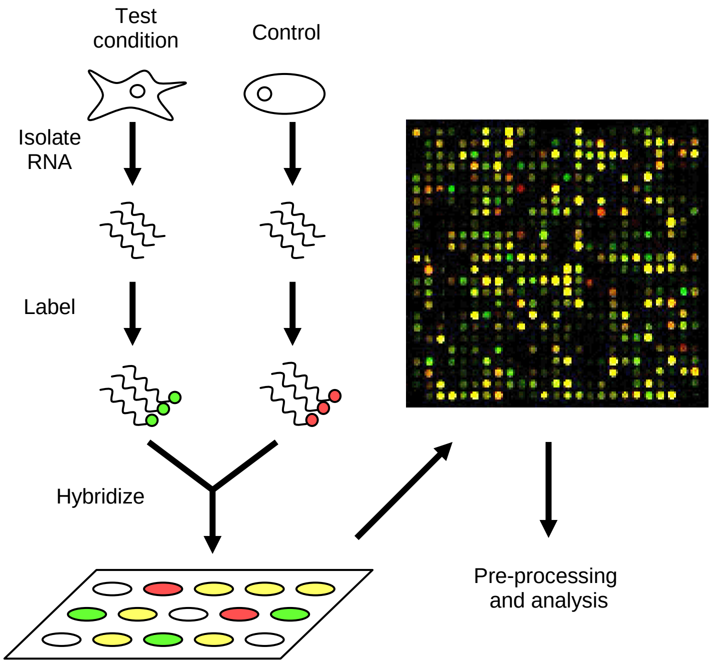


Figure 3: Two-color microarray experimental design

each sample can be inferred from the intensity of fluorescence in the red and green channels. High-density oligonucleotide arrays, such as the Affymetrix platforms, have generally superseded two-color arrays. Affymetrix arrays are single channel instruments, allowing for absolute quantification of expression [36]. With Affymetrix arrays, the abundance of target transcripts is related to intensity of fluorescent signal. Affymetrix platforms use 25 mer oligonucleotide probes to quantify gene expression. Multiple probes are targeted to specific sequences along the coding region of the genes represented on the array. Probes are combined into probe sets consisting of between 11-20 probes in order to improve the sensitivity of the assay. Mismatch probes, designed with one nucleotide base change from the perfect match probe, are intended to quantify the extent of non-specific hybridization.

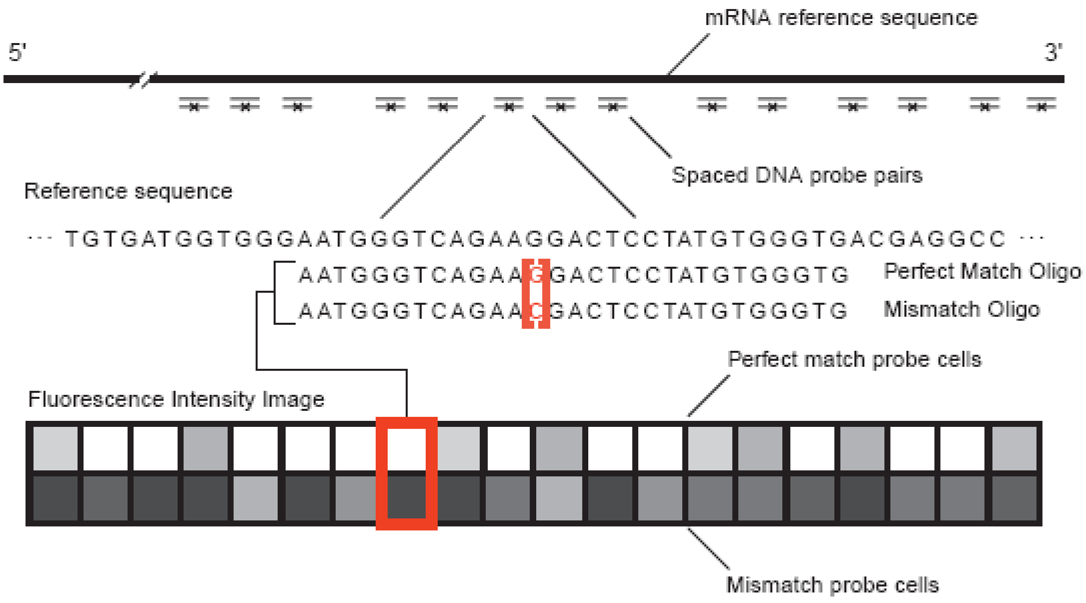

mRNA reference sequence
5'
3'
Spaced DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···
AATGGGTCAGAAGGACTCCTATGTGGGTG
Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTGGGTG
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 4: Probe sets on Affymetrix gene expression microarrays: Probe pairs are designed to hybridize to different segments along the coding region of the target gene. Pairs consist of a perfect match and mismatch probe which differ at a single base pair. [36].

In addition to gene expression platforms, Affymetrix has developed gene mapping arrays, capable of genotyping SNPs distributed evenly across the genome. Gene mapping arrays incorporate 25 mer oligonuclotide perfect match and mismatch probes, similar to expression arrays. Perfect match probes are designed to hybridize to one of two potential sequence variations, represented by a single base substitution at the middle of the probe sequence [37]. Probes are also included to match both the sense and complimentary anti-sense DNA strands. Genotype calls at each SNP location result from the detection of the presence or absence of the two associated signals.

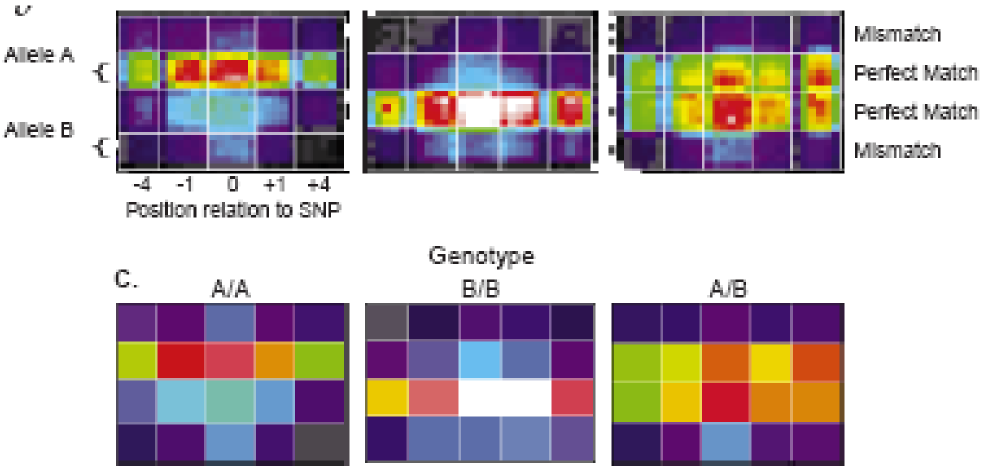


Figure 5: Fluorescent signal scanned from a gene mapping array: Rows two and three contain probes designed to hybridize to the A or B allele respectively. A heterozygous individual is identified when both A and B signals are present. [36]

A set of standards has been developed to promote the sharing of microarray data. The Minimal Information About a Microarray (MIAME) recommendations stipulate that researchers should provide open access to information regarding the experimental protocols and analytical methods used, sample and array annotations, and both raw intensity and processed expression data [38]. Open access standards ensure that the results derived from microarray analysis can be properly vetted by the scientific community and that the maximum benefit is gained from the data.

2.6 Biological databases

The rapid accumulation of biological knowledge has prompted the compilation of information into biological databases. Databases maintained by the National Center for Biological Information and the European Bioinformatics Institute provide useful resources for biological and medical research. Relevant gene information including the gene name, symbol, sequence, function and chromosomal location can all be found at these sources [39, 40]. Similar information on proteins is compiled [41]. Databases such as the Gene Expression Omnibus [42] and ArrayExpress [43] provide a repository for microarray expression datasets. Other databases have been developed with the goal of categorizing genes and proteins into logical groups. The Gene Ontology (GO) Database characterizes genes on the basis of three general categories: molecular function, cellular component and biological process [44]. Within each category, a nested vocabulary is established to classify genes with increasingly specific terms. The Kyoto Encylopedia of Genes and Genomes (KEGG) Database groups genes and gene products according to canonical signaling and metabolic pathways [45]. KEGG also maintains graphical representations of pathways for visualization. Compilation and logical organization of biological information in this manner facilitates biological research and accelerates the pace at which novel insights can be gained.

Chapter 3: Microarray Pre-Processing and Analysis

This chapter discusses methodologies for pre-processing and analysis of gene expression and gene mapping microarray data. Pre-processing corrects for technical variation through background correction, normalization and summarization of probe-level measurements. Analytical methods for microarray data identify differentially expressed genes, groups genes and samples according to common gene expression patterns, and detect gene expression patterns and gene sequence variations that are associated with phenotype.

3.1 Pre-Processing

Microarray technologies take advantage of the specific hybridization of oligonucleotides to complimentary sequences in order to quantify the abundance of specific transcripts or to characterize genetic variation at specific loci in the genome. Expression measures or genotypes are derived from the processing of fluorescent signals. In this process, there are many sources of obscuring variation, including non-specific hybridization, optical noise, reagent batch effects, microarray chip effects, and other stochastic differences in laboratory conditions [46]. Obscuring variation must be corrected before analysis so that the biological variation between the experimental conditions can be assessed. Pre-processing is used to correct for obscuring variation and derive adjusted expression values from the observed fluorescence intensity measures. Pre-processing generally consists of three steps: background correction, normalization, and summarization [47]. Background correction adjusts raw fluorescence intensities to remove signal originated from non-specific hybridization and optical noise. Normalization modifies background-corrected intensity values to remove obscuring variation and scale values such that they are comparable across arrays. Summarization is used to generate a single expression value for each probeset from the background-corrected, normalized intensity values of the constituent probes.

3.1.1 MAS 5.0 pre-processing algorithm

Affymetrix developed an algorithm (MAS5) for pre-processing gene expression arrays that treats each array separately[48]. The MAS5 background correction procedure calculates local background estimates across 16 rectangular regions of the chip mean as the lowest 2% of intensity values in each region. Probe intensities are adjusted by subtracting a weighted average of the local background estimates. Weights are dependent on the distance between the probe and the centroid of each region used for background estimation. Following this step, non-specific hybridization is accounted for by subtracting the mismatch (MM) probe intensity from the perfect match (PM) probe intensity. A separate procedure based on the average of MM and PM intensities is used to avoid negative values if MM is greater than PM. For normalization, a single array is chosen as a reference to which all of the remaining arrays are normalized against. A scale parameter is calculated by dividing the mean intensity of the reference array by the mean intensity of each non-reference array. Non-reference arrays are normalized by multiplying the intensity values by the corresponding scale parameter. The one-step Tukey’s biweight algorithm is used to generate expression values for each probeset. In this process, summarized probeset expression values are obtained by a weighted average of probeset intensity values in which the weights are defined on the basis of the uniform distance from the median intensity value. Analysis of benchmark datasets in which specific transcripts are spiked-in at certain concentration has shown that MAS5 has slightly lower precision than other algorithms [47].

3.1.2 Robust multi-array analysis pre-processing algorithm

Model-based algorithms, such as robust multi-array analysis (RMA) [47], are an alternative to MAS5 that use information across arrays to account for obscuring variation. It should be noted that RMA ignores the MM intensity values. Robust mutli-array analysis assumes that the observed fluorescent intensity consists of a normally distributed background component and an exponentially distributed signal component. A background estimate for each array is obtained by fitting the parametric model to the intensity values. Background is removed by subtracting the estimate from the perfect match intensity values. Background corrected intensity values are normalized using quantile normalization [49]. In quantile normalization, intensity values in each array are sorted in increasing order. Intensity values for each probeset are replaced by the average intensity values obtained by calculating the mean across arrays. At last, each array is restored to its original order prior to sorting. Application of quantile normalization results in equalization of the empirical distributions of each array. In order to summarize probe-level measures into probeset expression measures, RMA fits the background corrected, normalized and log base 2 transformed probe intensities to a linear model [47]. Model parameters are estimated using median polish to decrease the vulnerability to outliers. Analysis of spike-in benchmark datasets has validated that RMA produces expression measures with high accuracy and precision [47].

3.1.3 Reference robust multi-array analysis

Expression measures generated from multi-array pre-processing algorithms are highly dependent on the data used for normalization. This becomes a problem when adding new samples to an analysis. Reference RMA (RefRMA) [50] was developed to allow investigators to pre-process arrays using the same parameters generated from a reference dataset. Initially, RMA is run on a large, biologically diverse training dataset is to define a normalization vector through quantile normalization and a probe effect vector based on the probe affinities derived from the summarization step. These vectors can be applied in the pre-processing of newly collected arrays to calculate probe set expression measures that are comparable to the training dataset. Reference RMA can also be used to pre-process large datasets at a lower computational cost than RMA.

3.1.4 Bayesian robust linear model with mahalanobis distance algorithm for genotype calling

Algorithms used for pre-processing gene mapping arrays use information across chips and SNPs to evaluate the presence or absence of signal and make genotype calls. One such algorithm is the Bayesian robust linear model with Mahalanobis distance (B-RLMM) [51]. Similar to expression analysis, the goal of pre-processing is to convert measured fluorescent intensity values into an estimated genotype while correcting for obscuring variation. For each SNP represented on the array, summarization of probe intensity values results in two values, an A and a B signal, representing the two potential genetic variations. Normalization and summarization processes are similar to those used in the RMA algorithm. Briefly, probe intensity measures are quantile normalized, log transformed and median polish is used to estimate probe affinity effects. No background correction is necessary. Genotype calls are made on the basis of the summarized A and B signals. Transformations are used to represent the summarized signals in two-dimensional space. The contrast and size the signals are calculated as described below,

$$Contrast = (S_A - S_B)/(S_A + S_B)$$
$$Size = \log(S_A + S_B)$$

where S_A and S_B are equal to the A and B signals respectively. In this transformed space, the three potential genotypes (ie. homozygous A, heterozygous or homozygous B) are grouped into three clusters. A Bayesian procedure is used to define a unique set of clusters for each SNP. Briefly, a subset of SNPs is selected to generate an initial guess of the cluster distributions (prior). Next, signal values from each SNP are used to generate a specific estimate of cluster distributions. This specific estimate and the generic prior are used to generate a posterior estimate of the cluster distributions. Genotype calls are made by calculating the Mahalanobis distance between the transformed SNP signal values and the three clusters. The genotype corresponds to the minimum of these distances. Use of the B-RLMM algorithm produces a greater than 98% accurate call rate on reference samples derived from the HapMap project [51].

3.2 Gene expression microarray analysis

3.2.1 Differential expression analysis

Identification of differentially expressed genes is a common goal in microarray analysis. Given microarray data from samples in a number of classes, differentially expressed

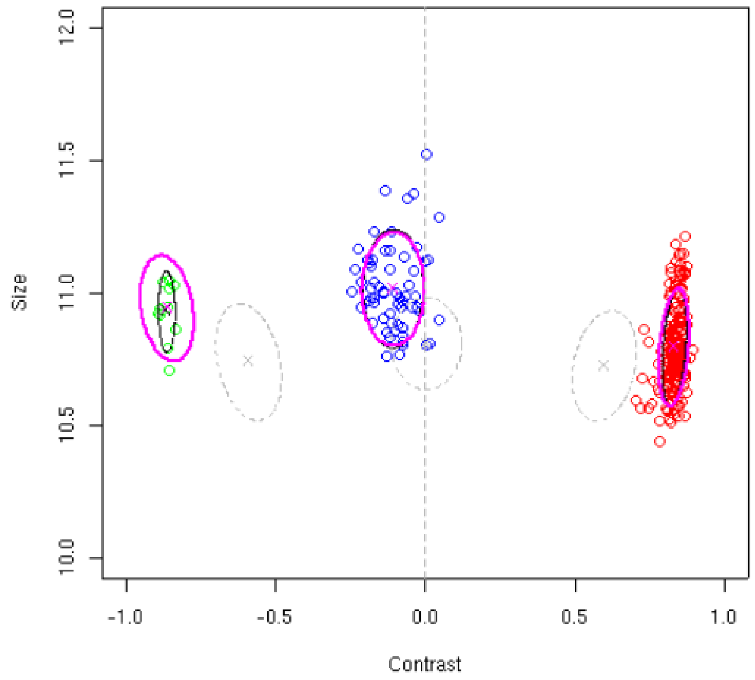


Figure 6: Genotype calling with B-RLMM – Genotypes represented as points in transformed space. Green points represent homozygous B genotypes, blue points represent heterozygous AB genotypes and red points represent homozygous A genotypes.

genes are expressed at a high level in condition A and a low level in condition B or vice versa. Initially, differentially expressed genes were identified using fold-change [52]. However, fold-change is less than optimal because it neglects the variance in expression measures. Statistical tests such as the Students t-test [53], anaylsis of variance (ANOVA) [54], or the Mann-Whitney test [55] have largely replaced fold-change in differential expression analysis. Parametric tests use the class-specific mean and variance to evaluate whether there is enough evidence to reject the null hypothesis of no differential expression. The Students t-test for example calculates a t-statistic as the difference in condition-specific means divided by an estimate of the pooled standard deviation [53]. T-statistics much higher or lower than zero indicate differential expression. A p-value can be calculated from the t-statistic to define the probability that the result could have been obtained by chance. Typically, p-values less than 0.05 indicate that the null hypothesis can be rejected. In expression analysis, the p-value must be adjusted to take into account the number of hypotheses tested (e.g. 5% of ~20000 genes tested is ~1000 false positive rejections of the null hypothesis). Multiple testing procedures such as the Bonferonni correction [56] or calculation of the false discovery rate [57] are used for this purpose. The Bonferonni correction simply divides the original significance level by the number of genes tested [56]. This correction is highly conservative and often results in no differentially expressed genes. The false discovery rate estimates the proportion of differentially expressed genes that are false positives [57]. An approximation of the false discovery rate can be obtained by permuting the class of the samples, calculating the test statistic and finding the number of genes that pass the significance threshold based on the permuted data [53]. Once a set of differentially expressed genes have been identified, new hypotheses regarding the active biological processes in each class of samples can be generated.

3.2.2 Unsupervised classification

Unsupervised classification methods utilize distance metrics to identify genes or samples with similar expression patterns. A gene expression profile can be defined as the vector of gene expression values across all samples or the vector of expression for all genes in a given sample. Genes with similar expression profiles tend to be co-regulated or involved in common cellular mechanisms [58]. With this observation, unsupervised classification methods have been used to infer the function of poorly characterized genes. In cancer studies, samples with similar expression profiles tend to have similar clinical characteristics. Clustering has been used to define cancer subtypes for prognostic purposes and selection of therapeutic strategies [59, 60].   Hierarchical clustering is an example of an unsupervised classification method. In hierarchical clustering, a nested tree-like structure is created by grouping the two most similar expression profiles in an iterative fashion [61]. Model-based clustering is an adaptation of unsupervised clustering methods that can be used to determine the confidence in cluster membership [62]. In model-based clustering, clusters are defined as multivariate normal distributions. Classification is based on fitting cluster-specific distributions to the data using either expectation maximization [63-65] or Bayesian methods [62, 66, 67]. Unsupervised classification methods are well suited for class-discovery in which the underlying structure of the dataset is unknown.

3.2.3 Supervised classification

Supervised classification methods use statistical hypothesis tests to identify significant genes and create a function capable of predicting the class of a new set of samples. This concept is similar to machine learning and consists of three steps: feature selection, classifier specification and evaluation of the predictor on an independent set of samples. Golub et al. used supervised classification to classify leukemia samples into acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) [59]. Genes associated with the ALL versus AML class distinction were identified using differential expression analysis. Specifically, the signal to noise ratio defined as the difference between class-specific mean expression values over the sum of the standard deviation in expression values from each class. A set of n genes with signal to noise ratios farthest from zero were selected as informative genes for classification. The number of genes parameter n was optimized using cross-validation on the training data. Classification of test samples was based on a weighted voting scheme. For each gene, the vote is calculated as the average expression value of the training samples subtracted from the expression value of the test sample. The class of the test sample is calculated based on the linear combination of the votes of each gene weighted by the associated signal to noise ratio. Classifiers are evaluated on an independent set of testing samples in which the classification is known. It is important to maintain the independence of the training and testing data to avoid biasing the classifier [68]. Supervised classification methods have been used to classify cancer samples on the basis of biomarker expression, lymph node involvement and subtype [59, 68].

3.2.4 Functional enrichment analysis

Once an interesting gene set has been identified through differential expression analysis or other means, further analysis can be used to determine which cellular processes are over-represented. Functional enrichment analysis compares the interesting gene set to a reference set of genes (ie. all genes represented on a microarray) [69]. Information from the Gene Ontology [44] or Kyoto Encyclopedia of Genes and Genomes [45] can be used to identify genes with common function. The hypergeometric test can be used to assess the significance of enrichment. Functional enrichment analysis can be used to infer which biological processes are activated or repressed by upregulation or downregulation of the genes in the interesting gene set.

3.3 Genotype analysis

3.3.1 Linkage mapping

Linkage analysis of genomic sequence data is used to identify genetic loci that are co-transmitted more than expected by chance under independent inheritance [70]. Both parametric and model-free methods have been developed to assess linkage. Parametric linkage analysis depends on the estimation of the recombination fraction, or the probability of recombination between two genetic loci, based on the observed genotype of related individuals. The genetic model of disease, including the mode of inheritance, frequency of the disease allele and penetrance of the genotype, must also be defined [70]. Linkage is determined by the calculation of a logarithm of the odds (LOD) score by a likelihood ratio test comparing the likelihood of linkage between a putative disease locus and a set of mapped marker loci against the null hypothesis of independent inheritance [70]. An LOD score greater than 3 indicates significant linkage. Non-parametric methods operate independently of genetic models by comparing the number of shared alleles between affected sibling pairs with the expected value [70]. Analysis of the linkage between genetic markers has been used to map the relative location of markers or disease loci in the genome. Linkage analysis is often used as a first pass to identify regions of interest for follow-up studies.

3.3.2 Association studies

Association testing of genomic data identifies alleles that contribute to disease susceptibility by comparing the frequency of occurrence in subjects with disease versus unaffected control individuals [71]. Simple association studies analyze the distribution of alleles in case and control populations with 2 x 2 contingency tables. Significance of association can be determined with the chi-square distribution [71]. The case-control study design may give false positive results in the presence of population stratification. Allele frequencies vary between different sub-populations according to genetic history. This issue can be avoided by stratifying the subject population on the basis of ethnicity. Methods have also been developed to estimate the population structure and adjust the test statistics appropriately [72, 73]. Alternatively, family-based association tests can be used. Family-based association tests (FBATs) avoid the confounding effect of population structure by evaluating the test statistic within families [74]. The FBAT statistic compares the genotype of affected and unaffected offspring to the expected value derived from parental genotypes under a Mendelian inheritance model. Signficance can be determined by comparing the magnitude of the test statistic against the normal distribution. Due to linkage disequilibrium, associated markers are considered to be in close proximity to the susceptibility polymorphism. In gene coding regions, analysis of the sequence and structure of homologous proteins can be used to differentiate between disease-related and disease-causing loci [75, 76].

Chapter 4: Prediction potential of candidate biomarker sets

4.1 Summary

This chapter discusses the identification and validation of candidate biomarker sets associated with molecular subtype and/or prognosis in multiple cancer types by the analysis of gene expression microarray data. Independently derived expression profiles of the same biological condition often have few genes in common. In this study, we created populations of expression profiles from publicly available microarray datasets of cancer (breast, lymphoma and renal) samples linked to clinical information with an iterative machine learning algorithm. ROC curves were used to assess the accuracy of each profile for classification. We compared the accuracy of profiles correlated with molecular phenotype against profiles correlated with relapse-free status. In addition, profiles identified with supervised univariate feature selection algorithms were compared to profiles selected randomly from a) all genes on the microarray platform (random selection) and b) a list of known disease-related genes (a priori selection). We also determined the relevance of expression profiles on test arrays from independent datasets, measured on either the same or different microarray platforms. Highly discriminative expression profiles were produced on both simulated gene expression data and expression data from breast cancer and lymphoma datasets on the basis of ER and BCL-6 expression, respectively. Use of relapse-free status to identify profiles for prognosis prediction resulted in poorly discriminative expression profiles. Supervised feature selection resulted in more accurate classifications than random or a priori selection, however, the difference in prediction error decreased as the number of features increased. These results held when expression profiles were applied across datasets to

samples profiled on the same microarray platform. Results suggest that many gene sets predict molecular phenotypes accurately. Given this, expression profiles identified using different training datasets should be expected to show little agreement. In addition, we demonstrate the difficulty in predicting relapse directly from microarray data using supervised machine learning approaches.

4.2 Background

Clinically validated biomarkers are highly valued in cancer pathology for diagnostic and prognostic purposes. Biomarker sets are also used in clinical trials as early indicators of drug efficacy and toxicity. Molecular profiling technologies have the potential to enable high-throughput candidate biomarker identification. Use of oligonucleotide or spotted cDNA microarrays allows for the quantification of the mRNA concentration of thousands of gene products simultaneously. Although measurement of the entire proteome is not yet possible, advances in mass spectrometry and chromatography provide similar capabilities at the protein level. Molecular profiling approaches have been applied towards the study of chronic diseases, including muscular dystrophy [77], diabetes [78], arthritis [79], cardiovascular disease [80] and cancer [59, 81-84]. Microarray studies in which the class or phenotype (e.g. health vs. disease, responders vs. non-responders, etc.) of all samples is known can be used to identify discriminative features (i.e. gene expression profiles) that are statistically associated with class distinction [59, 82-84]. These features can be used as potential biomarker sets to determine the phenotype of new samples and guide therapy appropriately.

Detection of candidate biomarkers from high-dimensional molecular datasets entails separation of signal from noise. As such, techniques adapted from signal processing and machine learning can be applied. The goal of machine learning is to reliably predict the class, or phenotype, of a new sample given only a set of measured input variables. The definition of a function that equates input variables to response is called supervised learning. In general, supervised learning consists of three steps: feature selection, decision rule specification and estimation of generalization error [68]. Feature selection is the identification of informative features from noisy or uncorrelated features in the dataset. Decision rule specification involves selection of a classification algorithm and definition of algorithm parameters by cross-validation [68, 85]. Feature selection and decision rule specification produce a classifier through the use of cross-validation on training data. In this process, there is a risk of overfitting the training data, in which the classifier is trained to recognize noise and not class distinction. The estimation of generalization error, or the misclassification rate expected when the classifier is applied to new samples, can be used to investigate the likelihood of overfitting. An unbiased estimate of the generalization error can only be obtained from independent test data [68].

Feature selection is particularly important in gene expression profiling, in which the number of features (genes) is much larger than the number of observations (microarray data samples). Identification of discriminative features eases the process of data interpretation and communication, decreases computation time for training, and, in biomarker identification, enables the development of reliable clinical assays. Numerous feature selection algorithms can be found in the literature, most of which rank features in a univariate manner, sorting them on the basis of correlation with class distinction [53, 85, 86]. In molecular profiling studies, univariate methods are used more often than multivariate feature selection methods [87-89] due to their intrinsic simplicity and the higher computational cost of multivariate methods.

Application of supervised feature selection methods in microarray analysis identifies a set of genes whose expression profiles are most correlated with response. However, discriminative feature sets identified in multiple microarray studies of the same disease state or biological condition typically share few common genes [90-92], indicating perhaps that multiple gene subsets can be used as effective biomarker panels. Many genes cluster into similar expression profiles and may have similar roles in signaling or metabolic pathways. Variation between studies can also be partially attributed to biological variations between sample populations and technical variations, such as the microarray platform (cDNA vs. oligonucleotide), protocol and analytical techniques used [93, 94]. Moreover, selection of discriminative genes within a given dataset is dependent on the selection of training set arrays [90, 95-97].

Given the presence of multiple, generally exclusive gene sets related to disease states such as metastatic breast cancer, it is appropriate to ask whether feature selection identifies gene expression profiles that classify better than is expected by chance, i.e. better than randomly selected gene sets. It is also important to determine to what extent technical and biological variability between studies affects the generalization error of classifiers trained on expression profiles. In this study, we analyzed a multitude of publicly available microarray datasets consisting of expression data linked to clinical data for breast cancer, renal cancer, and lymphoma [60, 82, 98-102]. Decision rules, composed of features associated with response, were created using supervised, univariate feature selection algorithms [81, 85]. Our analysis considered multiple microarray technologies (Affymetrix, cDNA spotted arrays, cDNA oligonucleotide arrays), normalization, feature selection and classification methods. Our results point to the efficiency of gene sets randomly selected from known disease-related genes in the accurate classification of cancer samples according to molecular phenotypes. Results also point to the challenges of predicting relapse directly from microarray data annotated with clinical outcome information.

4.3 Materials and methods

4.3.1 Microarray datasets

Publicly available gene expression data for a multitude of cancer types (breast cancer, lymphoma, and renal cancer) was collected from the online repositories Gene Expression Omnibus (GEO) [42] and Stanford Microarray Database (SMD) [103] (Table 1). All datasets used in the study were linked to clinical data including outcome and were further restricted to exclude datasets with less than 100 samples. Expression datasets analyzed in this article included data from multiple platforms (Affymetrix, cDNA, Hu25K), allowing us to assess the platform dependence of our results. Typically, datasets were collected from population-based studies with no age/status restrictions. Two exceptions to this rule are as follows: 1) dataset GSE2034 was restricted to breast cancer patients with lymph-

Table 1 - Description of microarray datasets used in biomarker set analysis

Disease Type	Datasets	Platform	# of Arrays	Restrictions	Reference
Breast Cancer	GSE3494	HG-U133a	251	a	Miller et al. [99]
	GSE2034	HG-U133a	286	b, c	Wang et al. [101]
	NKI	Hu25K	295	d, e, f	Van de Vijver et al. [82]
	Sorlie	cDNA	121	a	Sorlie et al. [60]
Diffuse large B-Cell Lymphoma	Broad	HG-U133a	176	a	Monti et al. [100]
	GSE4475	HG-U133a	220	a	Hummel et al. [98]
Renal Carcinoma	Zhao	cDNA	177	a	Zhao et al. [102]

a. No restrictions
b. Lymph-node negative
c. No adjuvant therapy
d. < 5 cm in diameter
e. <= 52 years at diagnosis
f. No previous history of cancer

node negative disease and with no adjuvant therapy; and 2) dataset NKI was restricted to patients with tumors less than 5 cm in diameter, and under age 52 at diagnosis (Table 1).

Each microarray dataset was analyzed independently to evaluate the error in predicting relapse (or histological expression of a surrogate biomarker of relapse) using univariate feature selection compared to the error from biomarker sets chosen randomly from either the entire set of genes represented on the microarray (random) or a smaller set of experimentally validated cancer-associated genes (a priori). To this end, we used an iterative supervised, machine-learning approach, described below. For completeness, we tested the dependence of our approach on the use of different pre-processing, feature selection and classification algorithms and cross-validation schemes. The primary focus of our study was on breast cancer, where multiple datasets were available for analysis. Lymphoma and renal carcinoma datasets were used to assess the relevance of our conclusions in other disease states. All work described in this study was carried out using the R statistical environment [104] and was duplicated independently in Matlab unless otherwise noted.

4.3.2 Pre-processing microarray data

Microarray datasets were collected in raw format when available (GSE3494, GSE4475). Two pre-processing algorithms, Robust Multi-Array Analysis (RMA) [46] and MAS 5.0 [48], were applied to these datasets to determine the effect of pre-processing on downstream analysis. RMA was implemented with the Bioconductor package [105] in the R statistical environment [104]. MAS 5.0 was implemented with Array Express Lite. All other datasets were obtained in pre-processed form. The methods used for pre-processing in these cases are summarized briefly as follows. The Broad and GSE2034 datasets were pre-processed using MAS 5.0. In the GSE2034 dataset, only chips with an average signal intensity of greater than 40 and a background signal of less than 100 were included and probe sets were scaled to a target intensity of 600 [101]. Sorlie and Zhao datasets were obtained from the Stanford Microarray Database (SMD) [103] in log base 2 form. Spots flagged by the scanning software were not included. Missing values were imputed using a nearest neighbor algorithm [106]. Expression values in the NKI dataset were quantified by averaging the intensity across the Cy3 and Cy5 channels and subtracting a local background estimate [82]. Each channel was normalized to the mean intensity across genes.

4.3.3 Probe set annotation

Probe sets on all platforms were annotated using gene identifiers maintained by the National Center for Biotechnology Information (NCBI). Affymetrix probe sets were annotated using the hgu133a package in R. Stanford clone identifiers were annotated using the SOURCE database [107]. Stanford cDNA datasets consisted of samples processed on different generations of cDNA platforms. To obtain comparable data within each dataset, we limited the dataset to the clone identifiers represented on all generations. This step resulted in 8404 and 39414 clone identifiers for the Sorlie and Zhao datasets respectively. The NKI probe sets were annotated using Unigene cluster identifiers from Unigene build 158 [108]. Retired cluster identifiers were identified and re-annotated using records from Unigene. These identifiers are sometimes split into multiple clusters. In these cases, annotation was not possible. These probe sets were excluded from the analysis. By retaining only the probe sets that could be definitively annotated, we were left with 8069 probe sets in the NKI dataset for further analysis.

4.3.4 Mapping between probe sets and genes

A single probe set representing each gene was selected to correct for the varying redundancy of gene representation on microarray platforms. In each platform considered in this study, approximately 60% of genes were represented by a single probe set. Genes represented by multiple probe sets were dealt with in the following manner. For Affymetrix datasets, probe set suffixes were used to remove redundant probe sets. For the HG-U133a chip, probe sets are encoded with _at, _s_at and _x_at suffixes that describe the quality of probe design [109]. All _x_at probe sets (~10% of probe sets on the array) were excluded. For genes represented by an _at probe set and multiple _s_at probe sets, the _s_at probe sets were discarded. Approximately 20% of redundant probe sets could be dealt with in this manner. In cases in which a unique probe set could not be chosen by the suffix, the average expression value of the remaining probe sets was used. For the non-Affymetrix probe sets, a unique probe set was chosen by selecting the probe set with the highest variance across samples.

4.3.5 Feature selection

Microarray datasets were iteratively divided into learning sets (LS) and test sets (TS) to create a population of classifiers and determine their classification performance in a Monte Carlo cross-validation approach [110]. Two types of response variables were used to divide samples into groups of poor prognosis and good prognosis, either histological expression of biomarkers (ER status in breast cancer, BCL-6 status in lymphoma) or relapse-free survival, in which relapse is defined as disease recurrence or death from disease. Learning sets and test sets were selected by first dividing datasets by response variable and then randomly selecting equal proportions of arrays from each class. Two different partitions were used: 2/3 LS, 1/3 TS and 1/2 LS, 1/2 TS. Learning sets were used to select informative features and train the decision rule. Genes were selected from the LS in a supervised manner using a univariate feature selection algorithm [85]. Briefly, each gene was ranked by the ratio of between class sum of squares to within class sum of squares. High scoring genes have large between class variances and small within class variances and are therefore correlated with class distinction. A second method was used to determine if our results were sensitive to feature selection algorithms

[81]. In this second method, genes are ranked by the signal to noise ratio, namely the ratio of the difference in class-specific means to the sum of the class-specific standard deviations [81]. This is quite similar to the two-sample t-statistic. We use the term signal to noise ratio to maintain consistency with previous literature in the field. For comparison, genes were selected randomly from either the entire list of genes represented on the array (random), or a list of experimentally validated disease-related genes obtained from the Ingenuity Pathways Database [111] (a priori). All three feature sets (feature selected, a priori, and random) were used in downstream analyses.

4.3.6 Classification

Two classification algorithms, diagonal linear discriminant analysis (DLDA) and k-nearest neighbour (NN, k = 3), were used to generate decision rules on the basis of the LS data. The NN algorithm classifies test samples according to the class of the three closest samples in the training set using Euclidean distance [85]. DLDA is based on the maximum likelihood discriminant rule [85]. These relatively simple classifiers have been shown to give accurate classifications in the analysis of expression data and appear to perform as well as or better then more sophisticated algorithms, such as support vector machines, and resampling methods, such as bagging or boosting [85, 86].

4.3.7 Validation

To obtain an estimate of generalization error, decision rules were applied to the corresponding TS. The confidence (δ) with which each sample was classified was calculated as follows:

$$\delta = \frac{d_R}{(d_R + d_N)}$$

in which d_R and d_N are distances measured between the test sample and the centroid of the poor and good prognosis LS, respectively. Samples are classified as good prognosis if the score is greater than 0.5 and vice-versa.

With this methodology, the classification performance of the decision rules could be visualized and compared with the use of receiver operating characteristic (ROC) curves [112]. ROC curves plot sensitivity, or detection rate, (β) against 1-specificity, or false alarm rate (α).

$$\beta = \frac{TruePositives}{(TruePositives + FalseNegatives)}$$

$$\alpha = \frac{FalsePositives}{(FalsePositives + TrueNegatives)}$$

Classification performance was determined from the area under the curve (AUC). Accurate classifiers have high sensitivity across the range of specificity and therefore have a large AUC. ROC curves were generated for each classifier using the ROC package in R. The score δ was divided into thresholds and β and α were calculated at each cut-off point. The AUC of each classifier was then calculated using the sum of trapezoids method. The entire process of feature selection, decision rule specification and estimation of generalization error was repeated 100 times to determine the expected performance of each gene set on a randomly selected set of samples. Average ROC curves were calculated from the distribution of detection rates at given false alarm rates [113]. Empirical confidence intervals were obtained as the 97.5% and 2.5% quantiles of this distribution. The expected classification performance was quantified using a prediction error metric E defined as 1-AUC. A smaller E value corresponds to a more accurate classifier.

4.3.8 Simulated expression data

Simulated microarray datasets were generated to verify that the machine learning algorithm described above leads to accurate classification of well-separated data. Simulated expression data was created in the manner described by Bura and Pfeiffer [114]. Datasets consisted of 100 observations with 1000 variables each, corresponding to arrays and genes respectively. Half of the observations were labeled as class 1 and the remainder were labeled class 0. All data for class 0 samples were drawn from a multivariate normal distribution with mean 0 and a covariance matrix of Σ. Five percent of genes were simulated to be differentially expressed. For class 1 samples, differentially expressed genes were drawn from a mixture of two multivariate normal distributions with means 0 and 2 and covariance structure Σ. The mixing probability was 1/2. Non-differentially expressed genes were generated from the same distribution as class 0 samples. The covariance matrix $\Sigma = \sigma_{ij}$ was generated with a block structure with $\sigma_{ij} = 0.2$ for $| j\text{-}i | \leq 5$ and 0 otherwise to model gene co-regulation.

4.3.9 Statistical significance of molecula profile prediction

To determine the significance of the calculated prediction error metric E for molecular profile prediction in breast cancer and lymphoma, the machine learning algorithm was repeated 1000 times with permuted class labels. An empirical p-value was calculated as the fraction of decision rules based on permuted class labels that performed better than the expected classification performance, E (described above), of decision rules based on true class labels. Permutation processes give an estimate of the likelihood that the true E value could be obtained by chance alone and are frequently used in similar studies for this purpose [53, 115].

4.3.10 Independent validation

In addition to cross-validated generalization error, we determined the classification accuracy across datasets. To this end, decision rules trained on one dataset were tested in both the corresponding test subset and datasets obtained by other laboratories. For across platform comparisons (Affymetrix vs. Hu25K, Affymetrix vs. cDNA), probe sets were matched by annotation to Entrez Gene identifiers.

4.4 Results and Discussion

4.4.1 Simulated datasets confirm the performance of classification algorithms

Analysis of simulated gene expression datasets indicated the effectiveness of the feature selection and classification algorithms used in this study to predict binary endpoints. Simulated datasets consisting of 100 observations and 1000 features were designed to approximate a binary classification problem [114]. Expression values were drawn from a multivariate normal distribution with mean equal to 0. Differentially expressed genes were simulated from a mixture of the original distribution with a second multivariate normal distribution with mean equal to 2. Our computations, presented in Figure 7, produced highly discriminative decision rules on simulated expression data. Elimination of differential expression, simulated by generating all values from the same distribution, resulted in classifiers with poor classification performance. This indicates that our algorithm accurately classifies well-separated data and avoids over-fitting the training data (Figure 7).

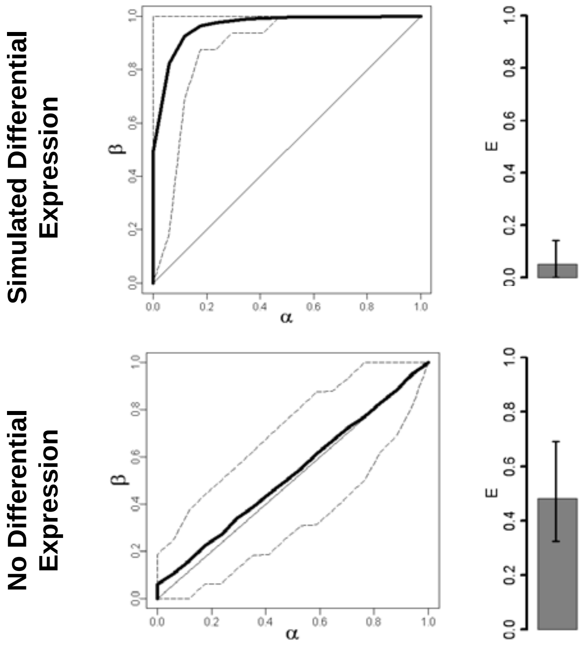


Figure 7 - Classification of Simulated Gene Expression Data. Receiver operating characteristic (ROC) curves showing classification performance of DLDA classifiers on simulated gene expression data. The symbols α and β are 1-specificity and sensitivity as described in the Methods section. Solid lines are average ROC curves over 100 iterations of training and test set selection. Dashed lines are empirical 95% confidence intervals. Bar plots give the mean 1-AUC (E) with error bars showing empirical 95% CIs.

4.4.2 Univariate feature selection is a poor predictor of relapse in breast, lymphoma and renal cancers

Computations on breast cancer microarray datasets from four independent cohorts of patients (GSE3494, GSE2034, NKI, Sorlie; Table 1) indicate the poor potential of univariate feature selection in predicting relapse-free survival. Figure 8 shows the classification error metric E (described in the methods section) as a function of the number of features used for classification. Columns 1 and 2 in this figure correspond to classification with respect to ER-status and relapse-free survival, respectively. Dark gray bars indicate univariate feature selection whereas light and medium gray bars correspond, respectively, to random selection from either the entire gene set or from an a priori gene set. Error bars indicate the variance over one hundred iterations of the machine learning algorithm. As the figure shows, decision rules trained on relapse-free status classify test samples with low accuracy. Analysis of diffuse large B-cell lymphoma (DLBCL) and conventional renal cell carcinoma (CRCC) datasets similarly yielded high errors in the prediction of relapse-free status (Figure 9). Survival time is a multi-factorial response variable with many potential confounding factors (e.g. lifestyle, age, etc.) that may affect gene expression. The influence of these confounding factors may result in tumor classes that are highly heterogeneous in regards to gene expression. These results indicate the difficulty in predicting relapse-free status in several forms of cancer from microarray data with the use of univariate feature selection.

3.4.3 Univariate feature selection as well as randomly chosen features from a priori knowledge set classifies microarray data according to molecular phenotype

In contrast, machine learning methods classified microarray datasets according to molecular phenotype with high accuracy (Figure 8). In analysis of breast cancer datasets, Figure 8 shows that decision rules trained on ER status classified test samples more accurately than decision rules trained on relapse-free status. These results agree with previous studies in that the expression profiles of many genes seem to be correlated with

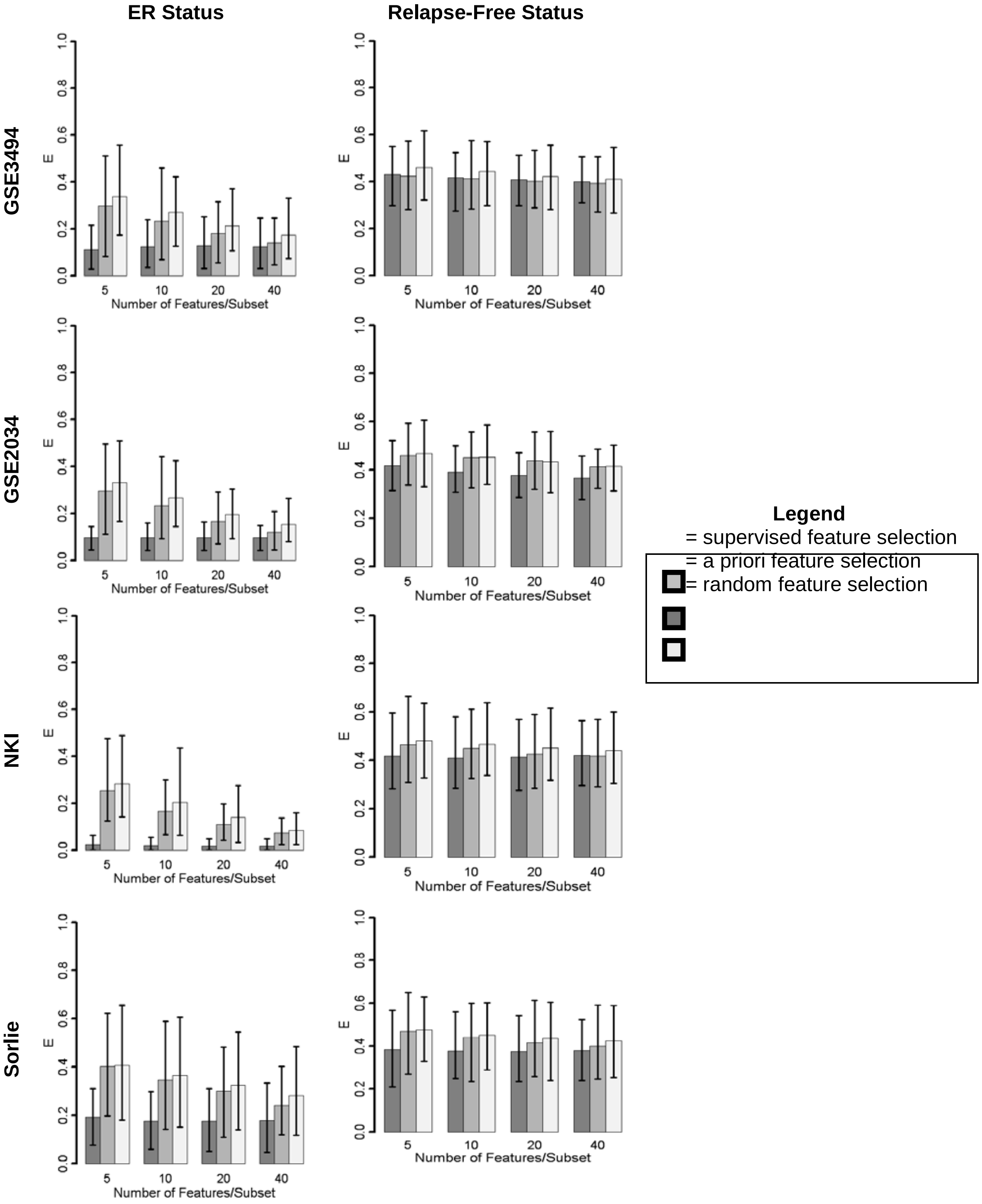

ER Status
Relapse-Free Status
GSE3494
GSE2034
NKI
Sorlie
E
0.0
0.2
0.4
0.6
0.8
1.0
5
10
20
40
Number of Features/Subset
Legend
= supervised feature selection
= a priori feature selection
= random feature selection

Figure 8 - Prediction error of DLDA classifiers trained and validated on breast cancer datasets. Column 1: Classifiers trained on ER-status. Column 2: Classifiers trained on relapse-free status. E is the mean 1-AUC of the corresponding set of ROC curves, calculated as described in the Methods section. Error bars show empirical 95% CIs.

ER status [83, 116]. Estrogen receptor is a hormone-activated transcription factor [117] and also participates in cellular signaling by heterodimerization with membrane-bound receptors such as the endothelial growth factor receptor [117]. Loss of estrogen receptor expression inhibits ER-responsive gene transcription and signaling in downstream pathways and therefore can be expected to affect the expression of downstream genes in a similar manner across tumors. Consistent with the analysis of breast cancer data, lymphoma datasets exhibited low errors in the prediction of BCL-6 status. BCL-6 is a zinc-finger protein that functions as a transcriptional repressor [118] and is expressed in germinal center B cells [119]. In DLBCL, BCL-6 expression, assessed by both immunohistochemistry and RT-PCR, has been associated with better survival in several studies [120, 121]. Univariate feature selection may be successful in predicting molecular phenotype due to the fact that expression profiles of many genes are correlated with changes in expression of these transcriptional modulators.

To determine whether gene sets identified with supervised feature selection are uniquely correlated with response, decision rules were generated both with and without supervised feature selection. In the absence of supervised feature selection, gene sets were drawn randomly from either the entire genechip (random selection) or a list of known disease-related genes (a priori selection)) (Figure 8). Random selection of subsets of *n* genes

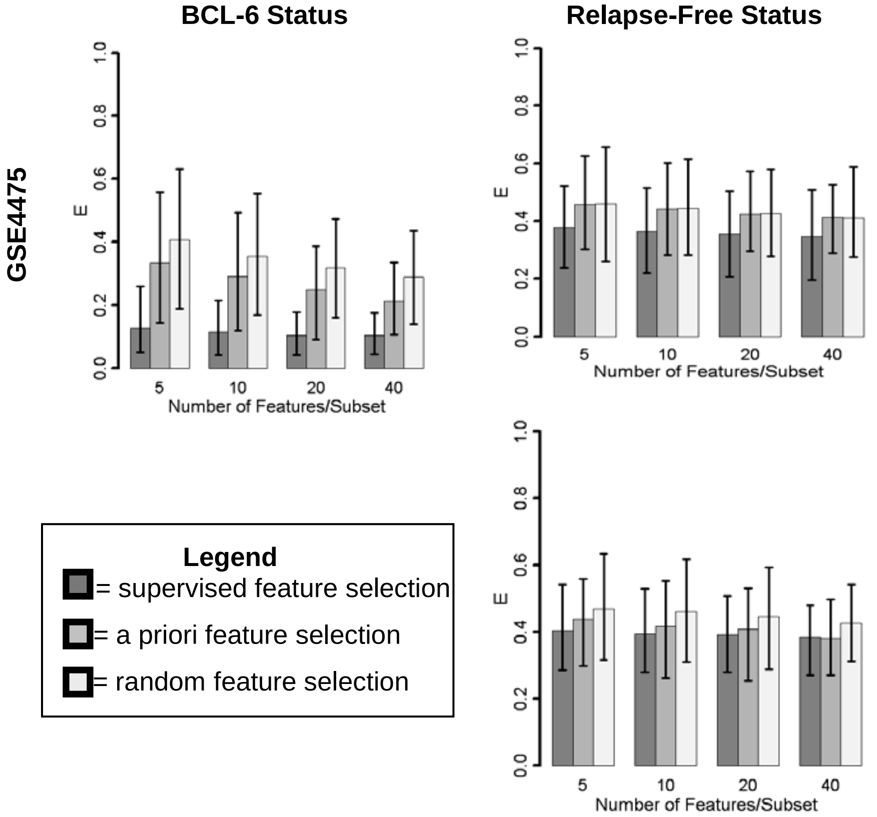


Figure 9 - Prediction error of DLDA classifiers trained and validated on diffuse large B-cell lymphoma and conventional renal cell carcinoma datasets. Column 1: Classifiers trained on BCL-6 status. Column 2: Classifiers trained on relapse-free status. Row 1: Diffuse large B-cell lymphoma. Row 2: Conventional renal cell carcinoma. E is the mean 1-AUC of the corresponding set of ROC curves, calculated as described in the Methods section. Error bars show empirical 95% CIs.

gives a baseline error rate expected for classification based on decision rules with *n* features. A priori selection provides a baseline error rate based on the known pathology. In Figure 2, we demonstrate that decision rules that incorporate supervised feature selection classify test samples more accurately than decision rules using a priori selection or random selection. However, in molecular phenotype prediction, the difference in prediction error decreases drastically as the number of features increases. This indicates that the power of univariate supervised feature selection methods lies in identifying small sets of discriminative features.

Exclusivity of predictive genesets has been demonstrated previously by investigating the classification potential of feature sets found down the list of genes ranked by their association with response [90]. Consistent with these previous observations, we demonstrate that randomly selected gene subsets classify molecular phenotype much better than the 50% error rate expected from random classification. In addition, limiting the feature space to genes that have demonstrated disease-relevance in the experimental setting improves classification performance of randomly selected gene sets. These results suggest that the presence of multiple, mostly exclusive biomarker sets identified from different studies [82, 83, 99, 101] can be partially attributed to the large number of combinations of discriminative feature sets [96].

4.4.4 Error in predicting relapse is insensitive to normalization and classification algorithms

Our computations indicate that classification error is only weakly dependent on normalization, feature selection, classification and training/testing partition. Breast cancer dataset GSE3494 was used to assess the effects of these classification parameters on predicted error. Results depicted in Figure 10 demonstrate that these parameters have little effect on the prediction of relapse-free survival, whereas pre-processing methodologies may have a small impact on the prediction error of ER status.

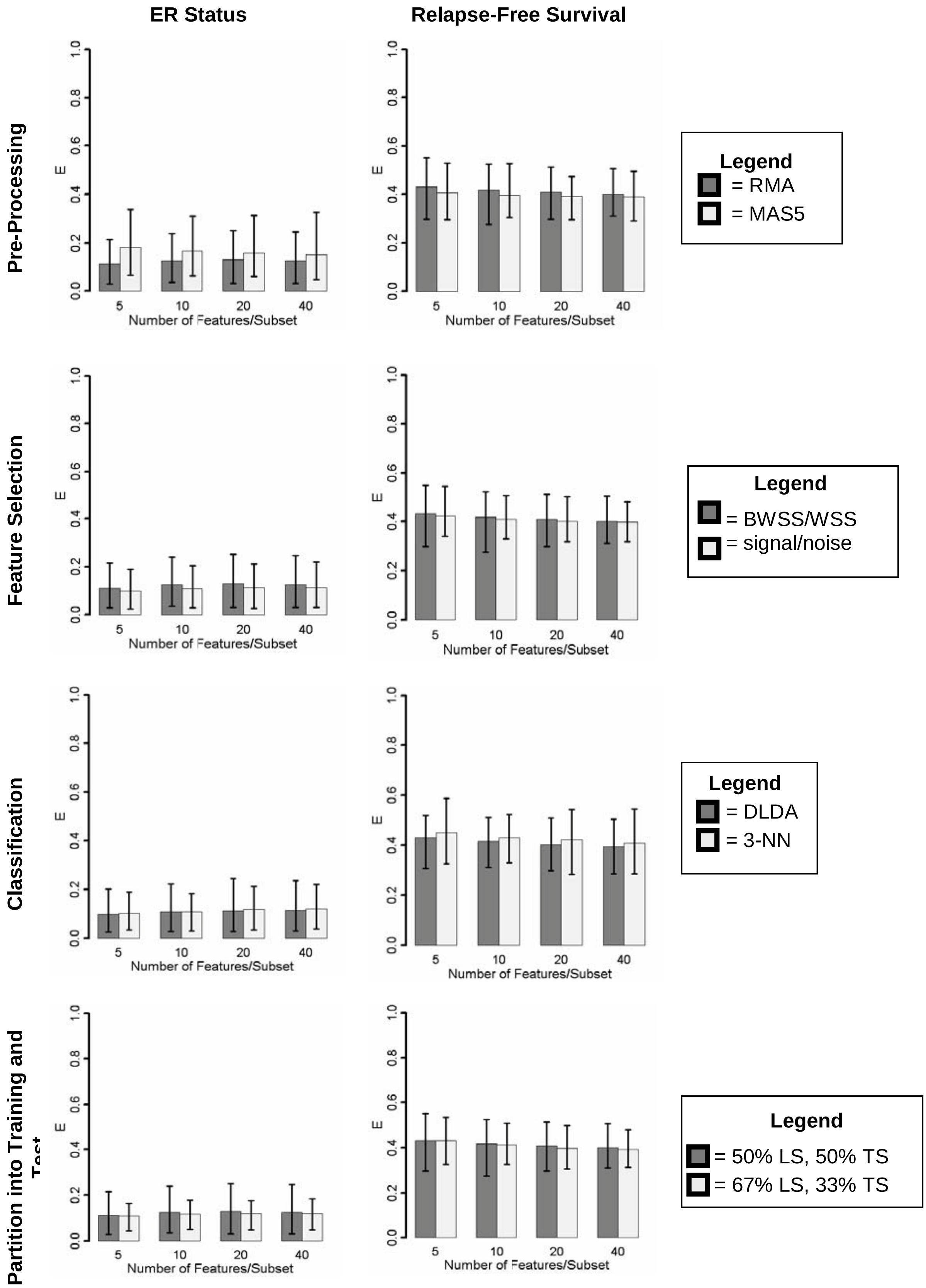
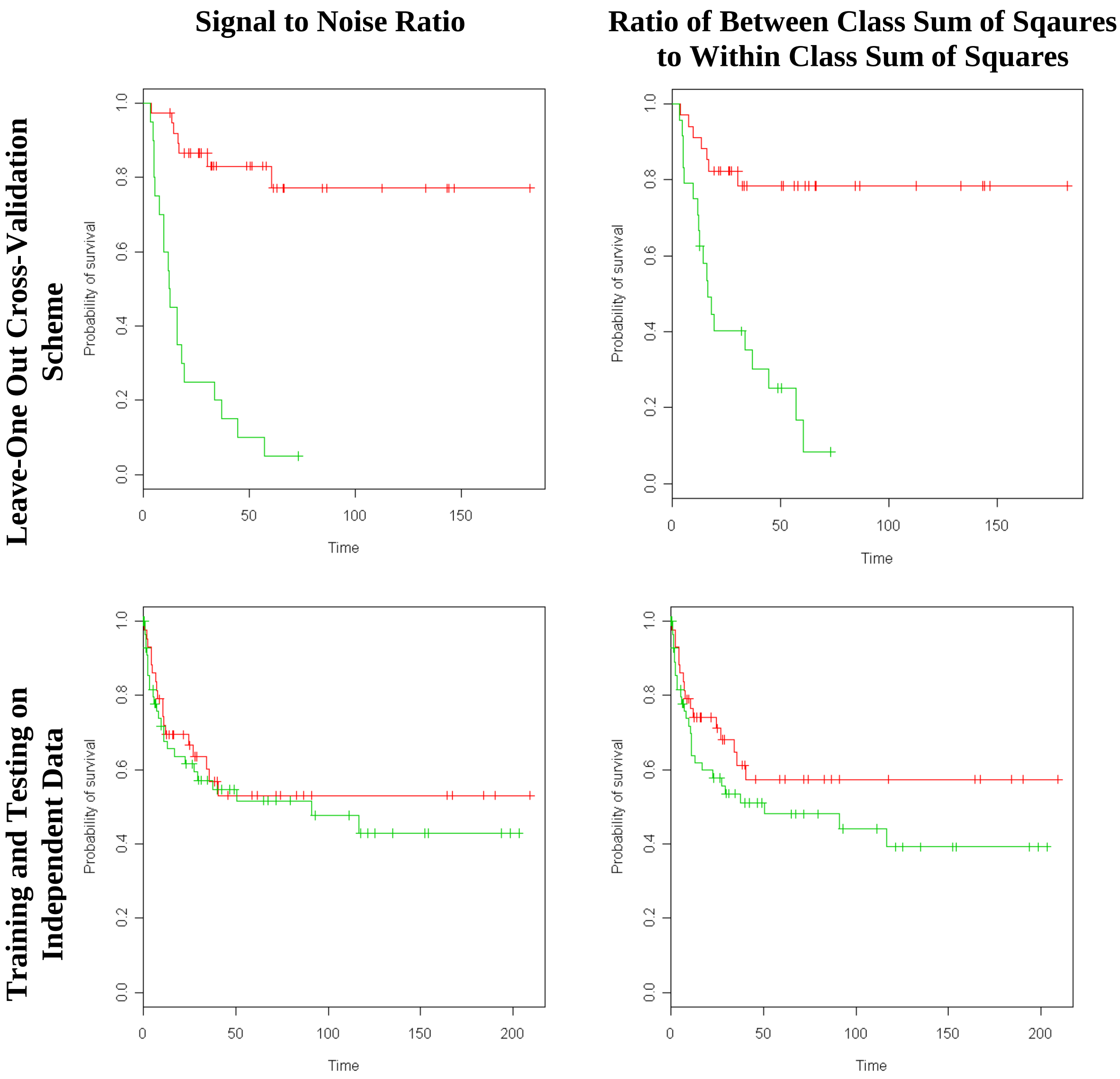

ER Status
Relapse-Free Survival
Pre-Processing
Feature Selection
Classification
Partition into Training and Test
E
0.0
0.2
0.4
0.6
0.8
1.0
5
10
20
40
Number of Features/Subset
Legend
■ = RMA
□ = MAS5
Legend
■ = BWSS/WSS
□ = signal/noise
Legend
■ = DLDA
□ = 3-NN
Legend
■ = 50% LS, 50% TS
□ = 67% LS, 33% TS

Figure 10 - Senstivity of classifiers to normalization and machine-learning parameters. Decision rules trained and validated on breast cancer dataset GSE3494 using supervised feature selection. Row 1: Expression values obtained using different pre-processing algorithms. Row 2: Different univariate feature selection methods. Row 3: Different classification schemes. Row 4: Different mode of partition into training and test data. E is the mean 1-AUC for the corresponding set of ROC curves, calculated as described in the Methods section. Error bars are empirical 95% CIs.

4.4.5 Leave-one-out cross-validation scheme may lead to overfitting

It has been shown that decision rules based on microarray data are capable of clearly differentiating tumors by outcome when all data is used for feature selection in a leave-one-out cross-validation scheme. Our findings validate previous results in the literature concerning, for example, the prediction of relapse in lymphoma [81]. In their study, Shipp et al. [81] used a machine learning procedure consisting of feature selection with the signal to noise ratio, classification by a weighted-voting scheme and leave one out cross-validation on a cohort of 58 lymphoma patients linked to clinical outcome. Importantly, the final geneset was selected from the consensus of all 58 leave-one out models of the data. Using Kaplan Meier analysis [122], Shipp et al. demonstrated a significant difference in survival between the classes predicted by machine learning. We replicated their calculations in this study using a larger microarray dataset (GSE4475, Table 1) and found similar results using both their and our methods of feature selection and classification (Figure 11, Row 1). Next, we divided the data in GSE4475 into a learning set (randomly selected set of 58 arrays) and test set (remaining 101 arrays) and computed Kaplan Meier survival curves. Results shown in row 2 of Figure 11 demonstrate the diminished capacity to identify groups of tumors with different survival rates when complete separation of training and testing sets is maintained in the



Signal to Noise Ratio

Ratio of Between Class Sum of Sqaures to Within Class Sum of Squares

Leave-One Out Cross-Validation Scheme

Training and Testing on Independent Data

Figure 11 - Kaplan-Meier plots of survival rates for predicted tumor classes with different feature selection/cross-validation methods. Classifiers trained on the basis of relapse-free status on diffuse large B-cell lymphoma dataset GSE4475. Column 1: Signal to noise ratio. Column 2: Ratio of between class to within class sum of squares. Row 1: Leave-one out cross-validation. All data used for training and testing. Row 2: Training and test sets selected randomly from the dataset. Training based on leave-one out cross-validation.

computations. If feature selection is included in the cross-validation procedure, such that features selected only from training data were applied to the test data, the difference in survival time between predicted classes decreases. These results suggest the possibility of overfitting in previously reported classifiers based on microarray data linked to clinical microarray data linked to clinical outcome.

4.4.6 Molecular phenotype prediction is maintained in across dataset cross-validation on the same microarray platform

Next, we tested whether prediction error calculated by within dataset cross-validation holds when decision rules trained on one dataset are applied to arrays from other datasets profiling similar populations. Within dataset cross-validation may be biased according to the degree of non-specific correlation between the training and test data. Non-specific correlation can be described as technical noise that arises in sample preparation, hybridization and scanning and results in higher correlation between data collected from the same lab compared to data collected in different labs [93]. To investigate this issue further, we used the Affymetrix dataset GSE3494 for developing decision rules for ER status prediction and applied these rules to arrays profiled on either the same (GSE2034) or different microarray platforms (NKI and Sorlie). There was no need to validate relapse prediction across datasets since our results showed poor prediction capacity even for within dataset cross-validation. Figure 12 illustrates the results of this analysis in the form of ellipses whose size and shape indicate the distribution of prediction errors. The column on the left (Column 1) corresponds to computations using univariate feature selection and the column on the right (Column 2) indicates results corresponding to random selection from an a priori dataset. The figure shows that the prediction error and its variance were much lower on test datasets profiled on the same platform (Figure 12, Row 1) in comparison to test datasets using different platforms (Figure 6, Rows 2 and 3). The same trend held true when the decision rule was based on feature selection from a random set

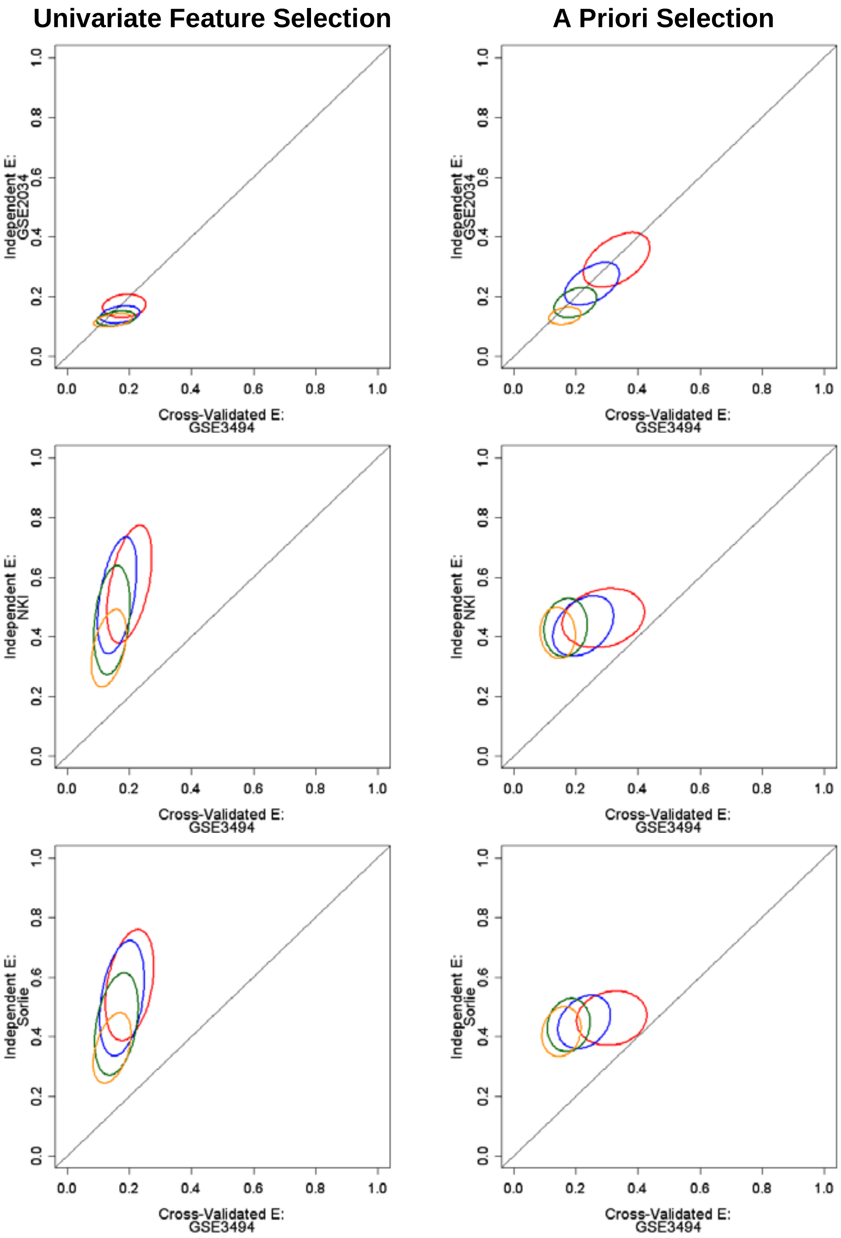


Figure 6. Prediction error of DLDA classifiers on breast cancer datasets by within-dataset and across-dataset cross-validation. Decision rules trained on ER-status. Ellipses are centered on the mean 1-AUC of the associated ROC curves. The major axis points in the direction of maximum variance. Lengths of the major and minor axes are proportional to the standard deviation of the data in each direction. Column 1: Prediction error of decision rules based on univariate ranking. Column 2: Prediction error of decision rules based on random selection of features from a subset with a priori disease relevance. 🟥 = 5 features/set, 🟦 = 10 features/set, 🟩 = 20 features/set, 🟧 = 40 features/set.

chosen with a priori knowledge (Figure 6, Column 2). These results suggest that decision rules obtained for classification do not accurately predict molecular phenotype in microarray data obtained using different platforms, possibly due to different strategies in probe design, or shortcomings in the matching of probes using probe set annotations [123]. Overall, these results demonstrate that bias resulting from non-specific correlation is negligible when samples are analyzed on the same platform. Results also validate the use of feature selection algorithms to identify small, discriminative feature sets that can be adapted for use in biomarker panels for identifying molecular phenotypes.

4.5 Conclusions

Biomarker sets derived from different gene expression microarray datasets for the purpose of predicting molecular phenotype or relapse in cancer contain very few common genes [91, 92]. In a typical microarray experiment, expression values of many genes are correlated with response [95, 96] and therefore, one could assume that multiple biomarker sets may accurately predict the classification of arrays into defined phenotypes. In this study, we used an iterative machine learning approach to determine the prediction potential of biomarker sets chosen using univariate feature selection from training sets selected randomly. On simulated gene expression data, this approach generated several highly discriminative decision rules. Similarly, multiple expression profiles capable of classifying tumors by molecular phenotype were identified in both breast cancer and DLBCL datasets.

We also compared the prediction error resulting from supervised feature selection vs. features selected randomly from either the entire set of genes represented on the microarray or an a priori defined subset of disease-relevant genes. Overall, univariate feature selection led to more accurate classification; however, the difference in prediction errors decreased as the number of features increased. Similar results were also observed in the application of decision rules to samples from other gene expression datasets profiled on the same microarray platform. From this, we conclude that the presence of multiple biomarker sets in the prediction of molecular phenotype arises from the large number of genes correlated with response.

In contrast, decision rules trained on the basis of relapse-free status classified samples with relatively high prediction errors in breast cancer, DLBCL and CRCC datasets. Specifically, prediction error was approximately 40% in all cases that were studied regardless of the method used for feature selection. Overall, these results indicate the difficulty of developing biomarker sets predictive of cancer relapse using a single microarray dataset. Our results do not apply to meta-analytical approaches, in which cancer relapse predictions are obtained by integrating data from multiple microarray datasets prior to machine learning [124-126]. In addition, combined use of clinical information and gene expression data may result in decision rules with better accuracy in predicting relapse [127-129].

Chapter 5: Expression profiles of switch-like genes in classification of tissue types and infectious disease

5.1 Summary

This chapter describes classification of tissue type and infectious disease phenotypes on the basis of the expression of bimodal, or switch-like, genes. Compilation of gene expression microarray datasets across diverse biological phenotypes has led to the identification and annotation of genes with bimodal expression patterns in the mouse and human genome. Approximately fifteen percent of known human genes exhibit switch-like expression profiles. Additionally, the switch-like gene set is enriched with genes expressed in the extracellular space and cell membrane. Evaluation of switch-like genes in large-scale microarray datasets may provide further insight into the biological relevance of bimodal gene expression patterns. In addition, it is of interest to determine the potential of bimodal genes for class discovery and class prediction. Use of a model-based clustering algorithm accurately classified more than four hundred microarray samples into nineteen different tissue types on the basis of bimodal gene expression. The algorithm demonstrated similar accuracy in the classification of microarray data corresponding to hepatitis C, influenza, HIV-1 and malaria infection. Classification accuracy was exceptional even with class-specific sample sizes between ten and twenty arrays. A supervised classification algorithm, in which feature selection was restricted to switch-like genes, also recognized tissue-specific and infectious disease specific expression profiles in independent test datasets reserved for validation. Classification of simulated microarray data indicated the validity of our observations in a large number of circumstances. Moreover, determination of consistent “on” and “off” states of switch-like genes in various tissues and diseases allow for the identification of

activated/deactivated genes and pathways. Functional enrichment analysis demonstrated that activated switch-like genes in neural, skeletal muscle and cardiac muscle tissue tend to have tissue-specific roles. A majority of activated genes in infectious disease are involved in processes related to the immune response. Our results indicate that switch-like gene sets capture genome-wide signatures from microarray data in health and infectious disease. Furthermore, we provide evidence that bimodal genes are involved in temporally and spatially active mechanisms including tissue-specific functions and response of the immune system to invading pathogens.

5.2 Background

Gene expression is controlled over a wide range at the transcript level through complex interplay between epigenetic modifications, DNA regulatory proteins, and microRNA molecules [14, 130, 131]. Genome-wide screening of expression profiles has provided an expansive perspective on gene regulation in health and disease. Identification of constitutively expressed housekeeping genes has aided in the inference of sets of minimal processes required for basic cellular function [132, 133]. Similarly, we have identified and annotated genes with switch-like expression profiles at the transcript level in the mouse and human, using large microarray datasets of healthy tissue [134]. Genes with switch-like expression profiles represent fifteen percent of the human gene population. Classification of samples on the basis of bimodal or switch-like gene expression may give insight into temporally and spatially active mechanisms that contribute to phenotypic diversity. Given the variable expression of switch-like genes, they may also provide a good candidate gene set for the identification of clinically relevant expression signatures.

The high-dimensionality inherent in genome-wide quantification makes extracting meaningful biological information from gene expression datasets a difficult task. Early attempts at genome-wide expression analysis used unsupervised clustering methods to identify groups of genes or conditions with similar expression profiles [61, 135, 136]. Biological insight can be derived from the observation that functionally related or co-regulated genes often cluster together. Supervised classification methods require datasets in which the class of the samples is known in advance. Statistical hypothesis testing [53, 92] is used to identify groups of genes that exhibit changes in expression associated with class distinction. Significant genes can be used to build decision rules to predict the class of unseen samples [83, 84, 101]. Unsupervised classification is better suited for class discovery whereas supervised classification is tailored for class prediction. In both of these complimentary approaches, dimension reduction can lead to increased classification accuracy.

Many simple unsupervised learning algorithms rely on distance metrics to either partition profiles into distinct groups [137, 138] or build clusters from pair-wise distances in a nested, hierarchical fashion [61]. The optimal number of clusters must be defined heuristically or in advance and confidence in cluster membership is difficult to determine. Model-based clustering provides the necessary statistical framework to address these concerns while allowing for class discovery. In model-based clustering, it is assumed that similar expression profiles are generated as draws from a set of multivariate Gaussian random variables. Clusters are identified by fitting the parameters of the cluster-specific distributions to the data. Expectation-maximization [63-65] or Bayesian methods [66, 67, 139] are used for optimization. Estimation of the number of clusters as well as the incorporation of confidence in cluster membership is implicit in this process.

Methods such as unsupervised, supervised and model-based classification provide the means to evaluate switch-like gene expression patterns in high-dimensional datasets profiling diverse biological conditions. In this study, we used these methods to identify tissue and disease specific expression signatures composed of switch-like genes. For this purpose, we compiled two large-scale gene expression microarray datasets from publicly available resources.   The first dataset included samples spanning nineteen different tissue types from healthy donors. The second dataset included samples from donors with one of a number of infectious diseases (HIV infection, hepatitis C, influenza, and malaria).   Our results demonstrate that bimodal gene expression profiles provide tissue-specific identification of samples in a dataset of healthy tissues. In addition, classification of switch-like expression patterns identifies infectious disease types with high accuracy and further specifies the tissue from which the diseased sample was obtained. Moreover, the set of activated switch-like genes for various disease and tissue types provide biologically significant information about the molecular basis of phenotype distinction.

5.3 Materials and methods

5.3.1 Microarray Datasets

Microarray datasets used in this study were compiled from online public repositories, the Gene Expression Omnibus (GEO) [42] and the Array Express (AE) [43]. All datasets were profiled on the HGU133A or its recently expanded version, the HGU133plus2 platform. The datasets used in the study are shown in Table 2.

5.3.2 Normalization

Prior to normalization, datasets were filtered such that only the 22,277 probe sets common to both the HGU133A and HGU133plus2 platforms were retained. Reference robust multi-chip averaging (refRMA) [50] was used for normalization. RefRMA is an adaptation of the classic RMA approach [47] that is better suited for large datasets. Briefly, RMA background adjustment was applied to each array. Arrays were normalized by fitting probe level intensities for each chip to an empirical distribution obtained by applying quantile normalization to an 800-array training set [134]. Probe affinity effects were estimated by median polishing on the training set and used to adjust the normalized probe level measures. Following these steps, probe set expression values were derived from the median value of constituent probe level intensities.

5.3.3 Probeset Annotation

Probe sets were annotated using entrez gene ID, emsembl accession number, gene symbol, Gene Ontology terms [44] and KEGG pathways [45]. Gene identifiers and gene ontology terms were obtained from the HGU133plus2 annotation information on the

Table 2 – Microarray datasets used in the analysis of bimodal expression patterns

Tissue Phenotype Data		
Tissue	No. of Samples	Gene Expression Omnibus/Array Express Accn. #
Adipose	10	GSE3526
Adrenal	20	GSE3526, GSE8514, GSE2316
Brain	89	GSE3526, GSE7621, GSE7307, GSE2361, E_AFMX-11, E-TABM-20,
Colon	10	E-TABM-176, GSE8671, GSE9254, GSE9452
Epidermal	25	GSE1133, GSE2361, GSE3419, GSE3526, GSE7307
Heart	38	E_AFMX-11, E-MIMR-27, GSE1133, GSE2240, GSE2361, GSE3526, GSE3585, GSE7307
Kidney	10	E_AFMX-11, GSE2004, GSE2361, GSE3526, GSE7392
Liver	10	E_AFMX-11, GSE2004, GSE3526, GSE6764
Lung	26	E-MEXP-231, GSE10072, GSE1133, GSE2361, GSE3526
Mammary	15	E-TABM-66, GSE2361, GSE3526, GSE7307, GSE7904
Muscle	64	GSE10760, GSE2328, GSE3526, GSE5110, GSE6798, GSE7307, GSE9103,
Ovary	10	GSE2361, GSE3526, GSE6008, GSE7307
Pancreas	6	GSE1133, GSE2361, GSE7307
Peripheral blood	12	GSE7462, GSE8608, GSE8668, GSE8762,GSE9692
Small intestine	7	GSE2361, GSE7307
Spleen	12	GSE2004, GSE2361, GSE3526, GSE7307
Stomach	10	GSE2361, GSE3526, GSE7307
Testis	38	E_AFMX-11, GSE1133, GSE2361, GSE3218, GSE3526, GSE7307, GSE7808
Thymus	5	GSE1133, GSE2361, GSE7307
Infectious Disease		
Disease	No. of Samples	Gene Expression Omnibus/Array Express Accn. #
Hepititis C	147	GSE11190, GSE7123
HIV	41	GSE6740, GSE9927
Influenza A	28	GSE6269
Malaria	15	GSE5418

Affymetrix website in March 2008. KEGG pathway annotations were obtained from the KEGG ftp site on April 28th, 2008.

5.3.4 Identification of bimodal genes

Bimodal genes were identified in expression data of healthy tissues [134] using a statistical method previously applied to detecting bimodality in blood glucose concentrations [140, 141]. For each gene, we tested the alternative hypothesis that the expression distribution fits a two-component Gaussian mixture model versus the null hypothesis that expression follows a single normal distribution. Identification of bimodality can be confounded in the presence of skew normal distributions. To correct for skewness, expression values were adjusted using the box-cox transformation [142]. Parameters of the two-component mixture model were fit using expectation maximization [143]. Parameters of the single normal distribution were estimated from gene-specific sample means and standard deviations. The log-likelihood ratio test statistic $-2\log\lambda$ was used to reject the null hypothesis. P-values were generated by evaluating the chi-square distribution with six degrees of freedom at the values of the test statistic. Genes with p-values less than 0.001 were selected as candidate bimodal genes. This subset of genes was further reduced by restricting the standardized area of intersection between the distributions of the component Gaussians [144]. We evaluated several increasingly stringent restrictions on the standardized area of intersection (<=0.1 and <=0.01). These thresholds produced sets of 1265 and 293 bimodal genes, respectively.

5.3.5 Identification of "on" genes in brain, skeletal muscle, cardiac muscle, lung and infectious disease phenotypes

Bimodal gene expression values were binarized by defining a gene-specific threshold at the intersection of the probability density functions of the two-component mixture models [144]. Expression values above this threshold are described as "high" or "on". Bimodal genes in the "on" state were identified using the Bernoulli process [144]. Briefly, each observation or sample was modeled as an independent trial. Success was defined as expression in the "on" mode. P-values were calculated from the binomial distribution with an equal probability of success and failure. A value of p<=0.01 indicates a significant association between bimodal gene expression and phenotype.

5.3.6 Functional Enrichment

Gene sets characterized by KEGG pathways and GO terms were analyzed to identify functional categories enriched with sets of bimodal genes biased to the "on" or "off" mode in healthy and disease phenotypes. We assessed the enrichment of functional gene sets by comparing the number of “on” or “off” genes observed in a particular functional group to the number expected by chance [69]. The hypergeometric test was used to assign significance to the enriched functional gene sets. P-values less than 0.001 were considered significant.

5.3.7 Distance-based clustering

Distance-based clustering algorithms implemented in the R statistical environment were used to classify tissue samples into groups with similar expression of bimodal genes. For completeness, we used both Kmeans and hierarchical clustering algorithms with Euclidean distance as a distance metric. Given a set of n observations defined in p-dimensional space (p = number of genes), the Kmeans algorithm partitions observations into K clusters by iteratively minimizing an objective function associated with cluster membership [145]. In our implementation, we ran Kmeans for ten iterations with ten different initial cluster centroid locations and retained the cluster partition associated with the minimal within-cluster sum of squares. Hierarchical clustering builds a dendrogram from the pair-wise distance between observations. We used complete linkage to define the distance between clusters and observations. A single cluster solution was obtained from the dendrogram by cutting the tree at a level which produced the desired number of clusters. In both of these algorithms, the data-driven optimal number of clusters was determined using the gap statistic, as described below.

5.3.8 Definition of the number of clusters in distance-based clustering

With the use of distance-based clustering in class-discovery problems, the optimal number of clusters $\hat{K}$ must be estimated from the data. We used the gap statistic [146] to test the null hypothesis that $\hat{K}$ = 1 i.e. no clusters. The optimal number of clusters was determined by comparing the within-cluster sum of squares to its expected value under a reference null distribution. The reference distribution was generated from a uniform distribution aligned with the principal components of the data as described by Tibshirani et al. Expression data was clustered into *k* groups (*k* = 1,2,...25) using either Kmeans or hierarchical clustering as described above. A set of *B* reference datasets were generated by drawing samples from the reference distribution and clustered in the same manner. The gap statistic was calculated as:

$$Gap_k = (1/B)\sum_b \log(W_{kb}^*) - \log(W_k)$$

in which W_{kb}^*, (*b*=1,2,...*B* and *k* = 1,2,...25) and W_k are within-cluster sums of squares of the reference and observed datasets respectively. The estimated number of clusters $\hat{K}$ is the smallest value *k* at which:

$$Gap_k \geq Gap_{k+1} - s_{k+1}$$

$$s_k = sd_k \sqrt{(1+1/B)}$$

and sd_k is the standard deviation of $\log(W_{kb}*)$.

5.3.9 Model-based subspace clustering

A model-based clustering algorithm [147], developed for the analysis of comparative genomic hybridization data, was used to cluster tissue samples on the basis of bimodal gene expression. In this approach, clusters are identified by finding an optimal partition of samples into K groups defined by cluster-specific multivariate Gaussian distributions. It is assumed that clusters can be differentiated by shifts in the mean expression values for a subset of genes. In the Hoff study, each sample is modeled as follows:

$$y_i = \mu + r_i \times \delta_i + \varepsilon_i$$

in which μ is a vector of mean expression values over all samples, $r_i \in (0,1)^m$ and indicates the relevant genes, δ_i is a vector of mean shifts and ε_i is a vector of the variance in expression values. Cluster-specific parameters $\Theta = (r_i,\delta i)$ are sampled from a baseline distribution f_0 in a Polya urn scheme or chinese restaurant process as described by Hoff:

sample $\Theta_1 \sim f_0$

sample $\Theta_n \sim \alpha/(\alpha+n-1)f_0 + (n-1)/(\alpha+n-1)f_{n-1}$

where f_{n-1} is the empirical distribution of $\Theta_1, \ldots, \Theta n$ and a is a constant. This process potentially results in less than n unique draws from the baseline distribution and therefore naturally leads to clustering. Parameters of the model are fit from the data using a Gibbs sampling algorithm. We ran the model-based clustering algorithm [147] in the R statistical environment on 25 parallel Markov chains with 250 iterations each. We found that each chain quickly converged to equally likely, unique solutions, indicating a multi-modal posterior distribution. To obtain an approximation of the true posterior distribution, we took the average of the cluster partition with the highest log-likelihood from each chain as reported elsewhere [67, 139].

5.3.10 Pairwise posterior probabilities

Given a set of clusters obtained from Gibbs sampling, the probability that two observations belong to the same class is approximated by the proportion of clusters in which they are grouped together [139]. For each pair of samples, the pairwise posterior probability matrix was calculated as:

$$P_{ij} = \frac{\#\,of\ \ clusters\ in\ which\ c_i = c_j}{total\ \#\,of\ \ clusters}$$

in which c_i (i = 1,…, n samples) is a vector indicating which cluster sample i is assigned to. Although the pairwise posterior probability is a useful measure in itself, it does not provide a single cluster partition. For this purpose, a distance metric was defined from the pairwise posterior probabilities equal to $D_{ij} = 1 - P_{ij}$ [139]. A unique cluster partition can then be found using the complete linkage method, such that objects are grouped together when the pairwise distance between them is less than one.

5.3.11 Quantifying the agreement between observed clusters and known phenotype

In this study, clustering algorithms were applied to data in which the true class membership of all samples was known *a priori*. The Adjusted Rand Index (ARI) was used to measure the amount of agreement between the known and estimated class membership [62, 67]. Given two partitions of n observations U = (u1,...,uR) and V = (v1,...,vC), where U indicates the cluster partition and V indicates the true class, the Adjusted Rand Index can be calculated from the contingency table of the two partitions (Table 3). An element n_{ij} of the contingency table equals the number of observations in cluster i that are of class j. Row sums of the contingency table are equal to $n_{i.}$ and column sums are equal to $n_{.j}$.

Table 3: Contingency table comparing two partitions

	v_1	v_2	…	v_C	
U_1	N_{11}	N_{12}	…	N_{1C}	$n_{1.}$
U_2	N_{21}	N_{22}	…	N_{2C}	$n_{2.}$
…	…	…		…	…
u_R	n_{R1}	n_{R2}	…	n_{RC}	$n_{R.}$
	$n_{.1}$	$n_{.2}$	…	$n_{.C}$	$n_{..} = n$

With this notation, the Adjusted Rand Index is calculated by the formula below and takes a value of 1 when the two partitions agree completely and a value of 0 when the index equals its expected value (i.e. the partitions are no better than random).

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_{i}\binom{n_{i.}}{2}\sum_{j}\binom{n_{.j}}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i}\binom{n_{i.}}{2} + \sum_{j}\binom{n_{.j}}{2}\right] - \left[\sum_{i}\binom{n_{i.}}{2}\sum_{j}\binom{n_{.j}}{2}\right]\Big/2}$$

5.3.12 Supervised Classification

A multi-class supervised learning scheme was used to classify tissue samples on the basis of bimodal gene expression. To extend the supervised learning scheme to multiple class problems, we trained separate classifiers to identify tissue samples of each class vs. all others [148]. Results are based on 100 independent iterations of the following training and testing procedure. Prior to classification, datasets were divided into training and testing sets in a class-proportional manner such that two-thirds of the samples in each class were used for training and one-third for testing. For the *j*th classifier (j = 1,... , number of classes), training samples in class *j* were assigned to class 1. All other samples were assigned to class 0. Discriminative bimodal genes were identified from the training data according to the ratio of within class to between class sum of squares [85]. Diagonal linear discriminant analysis was used to define the distances between test sample *i* and samples in class 0 (d_{co}) and class 1 (d_{c1}), respectively [85]. A confidence measure, defined from 0 to 1, was calculated as $d_{co}/(d_{co}+d_{c1})$. Values close to 0/1 indicate low/high confidence that test sample *i* belongs to class *j*. Confidence measures are compared from each classifier and test sample *i* is assigned to the class associated with the highest confidence.

5.3.13 Simulated Data

Synthetic data was used to determine the effect of sample size, effect size and the number of informative genes on prediction accuracy in binary classification. *In silico* expression datasets consisted of 10, 20, 30, 50, or 100 observations/arrays and 1000 features/genes. Initially, a binary vector indicating the class membership of each observation was drawn from a binomial distribution B(n,0.5). A number of 5, 10, 20, 50, or 100 informative gene expression profiles were drawn from a pair of multivariate normal distributions $N_1(\mu_1, \Sigma)$ and $N_2(\mu_2, \Sigma)$ representing each class of observations. Non-informative expression values representing noise genes were drawn from a mixture of N_1 and N_2 with mixing probabilities of ½ from each distribution. A diagonal covariance matrix (Σ) was used to simulate independent expression values. Effect size was measured by a separation parameter defined for each gene, specifically the distance in class-specific means divided by the pooled variance. Three effect sizes (6, 2, 1) were investigated. We used logistic regression to generate the response variable that indicates class membership from the expression data. Regression coefficients associated with the informative genes were drawn from a uniform distribution U(0.1,1). By logistic regression, the probability that the *i*th observation is class 1 is given by π_i:

$$\pi_i = \frac{1}{1 + \exp\left(\beta_1 x_{1,i} + \ldots + \beta_M x_{M,i}\right)} \tag{7}$$

in which $\beta_1 \ldots \beta_M$ are the defined regression coefficients and $x_{1,i} \ldots x_{M,i}$ are the expression values of the informative genes in the *i*th observation. The simulated dataset was completed by drawing the response variable y_i on the basis of π_i ($y_i = 1$ iff $\pi_i > 0.5$). In specified exactly (i.e. the value of β), independent of the sample distribution of gene *j*.

5.4 Results

5.4.1 Model-based clustering accurately classifies tissue phenotypes on the basis of bimodal gene expression

A model-based classification algorithm [147] partitioned a set of 407 microarray data samples into bins specific to 19 different tissue types (Figure 13). Classification was based on expression of 1265 switch-like genes with bimodal gene expression patterns identified in human microarray data [134]. In model-based clustering, the number of clusters is optimized as part of the model-fitting process. Each instance of model-based clustering leads to slightly different results, with some tissue samples appearing in different clusters from run to run. The posterior distribution of model-based cluster solutions captures the uncertainty in clustering and is approximated by summarization of the most likely partitions visited by the algorithm [66, 139]. To avoid the label-switching problem in summarization, in which the number and label of clusters differs between runs, we calculated the posterior pairwise probability that each pair of samples clusters together [62]. Heat maps shown in Figure 13 depict the posterior pairwise probability matrix for each pair of samples. The color of element x_{ij} of the heat map indicates the number of partitions in which sample i and sample j are assigned to the same cluster, with yellow being the maximum and blue the minimum. Rows and columns of the heat map are organized to group samples of the same tissue type together. The overlaid grid shows the boundary between different phenotypes. The figure shows that model-based classification correctly grouped microarray samples into tissue-specific clusters, even for tissues with as few as five microarray samples.

Two distance-based clustering algorithms, Kmeans and hierarchical clustering, were also used to classify tissue phenotype for comparison. We determined the optimal number of groups prior to distance-based clustering using the gap statistic [146]. Both distance-based clustering algorithms identified brain-specific and muscle-specific clusters but failed to differentiate between tissues with smaller number of samples (Figure 13). Partitions generated by model-based clustering reflect tissue phenotype more closely, as indicated by the yellow regions along the diagonal of the heat map (Figure 13).

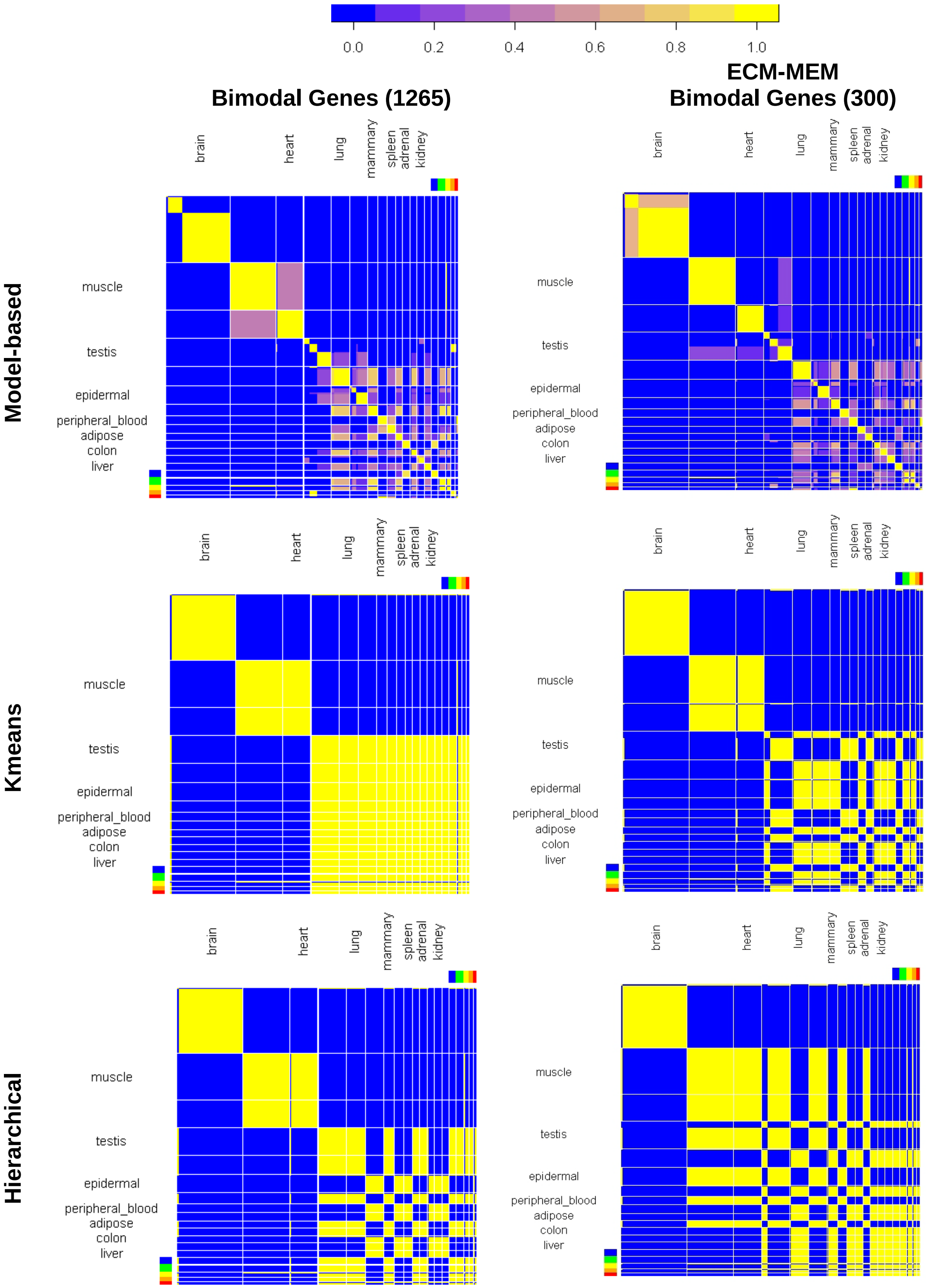

0.0
0.2
0.4
0.6
0.8
1.0
Bimodal Genes (1265)
ECM-MEM
Bimodal Genes (300)
Model-based
Kmeans
Hierarchical
brain
heart
lung
mammary
spleen
adrenal
kidney
muscle
testis
epidermal
peripheral_blood
adipose
colon
liver

Figure 13 - Model-based clustering of bimodal gene expression identifies cohesive clusters in 19 tissue types. Heat map representation of posterior pairwise probabilities for classification of tissue phenotype. Left column: classification with 1265 bimodal genes. Right column: classification with 300 bimodal genes translated into extracellular matrix or plasma membrane proteins. Top row: Model-based clustering identifies all tissues distinctly. Middle and bottom rows: Kmeans and hierarchical clustering classify samples into three/four tissue types: brain, cardiac and skeletal muscle and remaining tissues. Blue, green, yellow, orange and red regions of color bar indicate ovary, stomach, small intestine, pancreas and thymus tissue samples respectively. Tissues in the heat map were ordered according to decreasing sample size from left to right.

To further quantify our results, we used the Adjusted Rand Index (ARI) to evaluate how well the clustering algorithms extracted the class structure present in the tissue phenotype dataset. The Adjusted Rand Index measures the amount of agreement between the true class membership and the observed cluster partitions [62, 67]. ARI is equal to one at perfect agreement whereas it is equal to zero when the agreement is no better than expected by chance. Distance-based algorithms generate single partitions of the data that are suitable for analysis with the ARI. Prior to analysis of the model-based clusters, the posterior pairwise probability matrix was converted to a distance metric and a single partition was obtained via the complete linkage method [62]. As shown in Table 4, model-based clustering has significantly higher ARI values compared to distance-based clustering, especially for classification using a subset of bimodal genes in extracellular matrix and cell membrane GO categories. Consistent with the heat maps shown in Figure

Table 4 – Adjusted Rand Index compares observed partitions with true classification of samples in tissue phenotype data

	Kmeans	Hierarchical	Model-based
All bimodal genes	0.291	0.463	0.683
ECM/MEM genes	0.456	0.304	0.881

1, Table 2 shows that model-based clustering outperformed distance-based algorithms in unsupervised classification of tissue phenotypes.

5.4.2 Bimodal genes specific to extracellular matrix and membrane cell compartments improve model-based clustering of tissue phenotypes

Microarray samples profiling different tissue types were classified with a subset of bimodal genes whose products are expressed in the extracellular matrix (ECM) or on the plasma membrane (MEM). ECM and MEM genes are statistically enriched in the bimodal gene sets for the human [134]and the mouse[144]. Cell-cell or cell-ECM interactions, mediated through cell surface receptors, activate downstream transcriptional programs that regulate a diverse set of processes including growth, proliferation, apoptosis, and cell motility [149, 150]. A subset of ECM and membrane bound proteins are known to be tissue-specific and play crucial roles in the development and maintenance of tissue differentiation [151, 152]. Moreover, altered expression of ECM and MEM proteins has been linked to pathogenesis in muscular dystrophy, multiple sclerosis, and various cancers [153-155]. Using information obtained from the Gene Ontology database [44], 300 bimodal genes annotated with ECM or MEM terms were identified. Model-based clustering based on expression of ECM and MEM bimodal genes led to more accurate classification. Entirely separate clusters of skeletal and cardiac muscle were resolved. Other tissue phenotypes were also identified with higher accuracy as indicated by the color of off-diagonal elements in the heat map and the ARI (Figure 13, Table 4). Noting that the tissue-specific sample size in the microarray data ranged from 5 to 89 (Table 2), results with model-based classification indicate the strength of tissue-specific signatures in global gene expression and the ability of bimodal genes to capture such signatures. Results also indicate that a subset of bimodal genes whose products are positioned either on extracellular matrix or the cell membrane is sufficient to identify tissue-specificity in microarray data. Given the importance of ECM and MEM proteins in the regulation of cellular function, products of these genes may serve as candidate biomarkers or therapeutic targets in tissue-specific diseases.

5.4.3 Bimodal genes classify more accurately than randomly selected genes

Next, we clustered the tissue phenotype microarray datasets with randomly selected gene expression values. In previous work, we had demonstrated the power of randomly selected genes in supervised classification of molecular phenotypes in gene expression data of various cancers [156]. Random datasets consisted of expression values from 300 probe sets sampled without replacement from the total 22,277 probe sets analyzed. A total of ten random datasets were clustered and the posterior pairwise probability was calculated as described above. Both distance-based algorithms revealed a strong brain tissue expression signature, indicated by the observation that brain samples cluster together more often than with samples of other tissue types in random datasets (Figure 14). Kmeans clustering of random datasets also identified a strong muscle tissue signature (Figure 14). For comparison, we clustered the tissue phenotype data using a similar number of bimodal genes (293), identified by increasing the stringency of the

signature (Figure 14). For comparison, we clustered the tissue phenotype data using a similar number of bimodal genes (293), identified by increasing the stringency of the

tests used to detect bimodality. Classification of brain and muscle tissue was more accurate with the use of these bimodal genes than the randomly selected gene sets (Figure

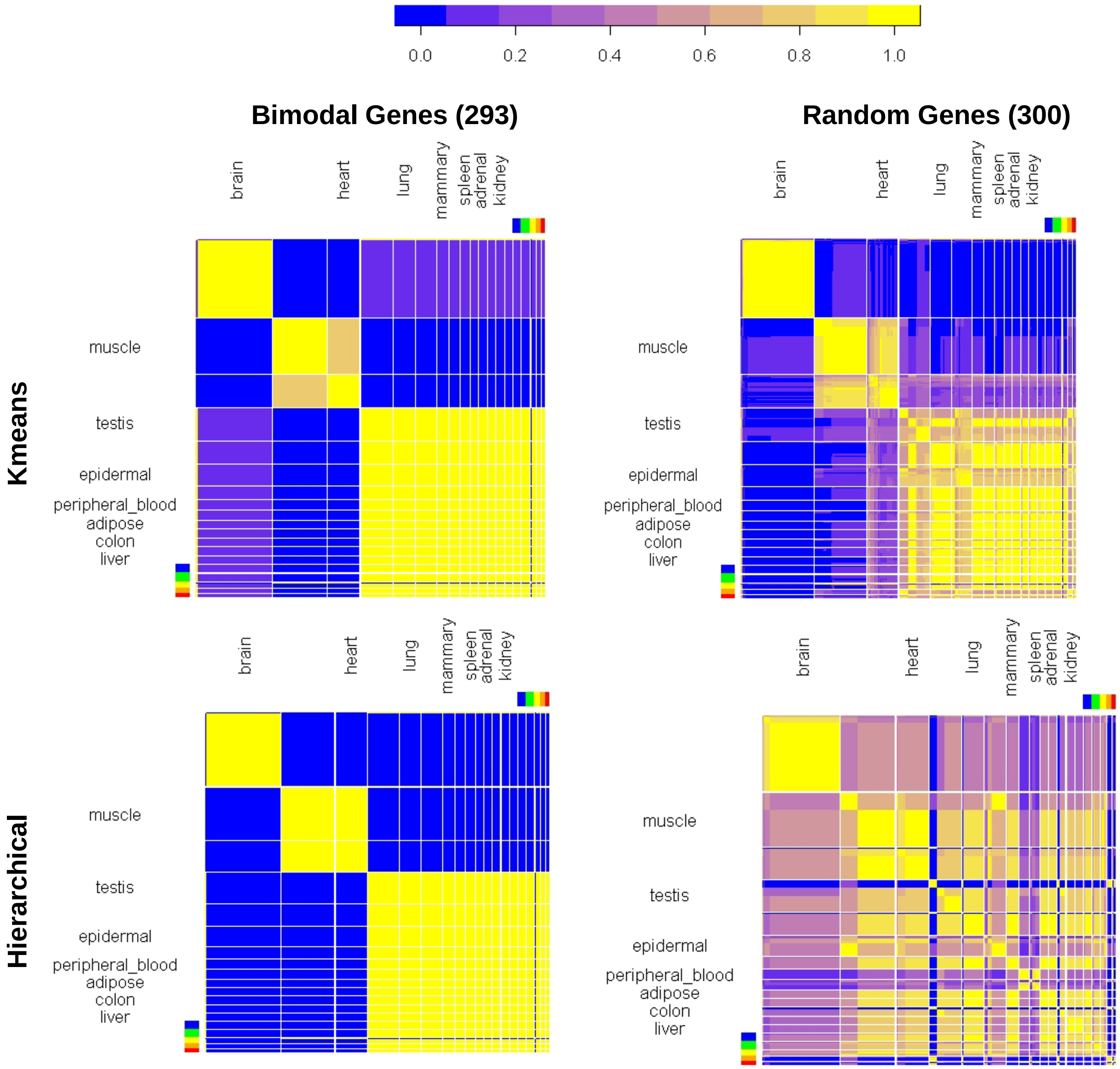


Figure 14 - Bimodal gene expression classifies tissue types more accurately than expression of randomly selected genes. Heat map representation of posterior pairwise probabilities for classification by bimodal and randomly selected genes. Top row: Kmeans clustering. Bottom row: hierarchical clustering. Left column: classification with 300 randomly selected genes on the microarray chip. Right column: classification with 293 bimodal genes (p-value<=0.001 and area of intersection <= 0.01). Blue, green, yellow, orange and red regions of the color bar indicate ovary, stomach, small intestine, pancreas and thymus tissue samples respectively.

14). Although there are strong brain and muscle tissue expression signatures in the data, bimodal genes appear to be enriched with more tissue-specific genes than expected by chance.

5.4.4 Enrichment analysis reveals tissue-specific functions of "on" genes in brain, skeletal muscle, cardiac muscle, and lung tissue

Functional enrichment identified gene sets related to tissue-specific function in sets of bimodal genes biased toward the "on" mode in a majority of samples of brain, skeletal muscle, cardiac muscle and lung tissue. Bimodal gene expression values were binarized into "on" and "off" modes prior to analysis as described in Ertel & Tozeren [144]. A gene by sample heat map (Figure 15) shows the mode of expression for all 1265 bimodal genes in 217 samples of brain, skeletal muscle, cardiac muscle and lung tissue.

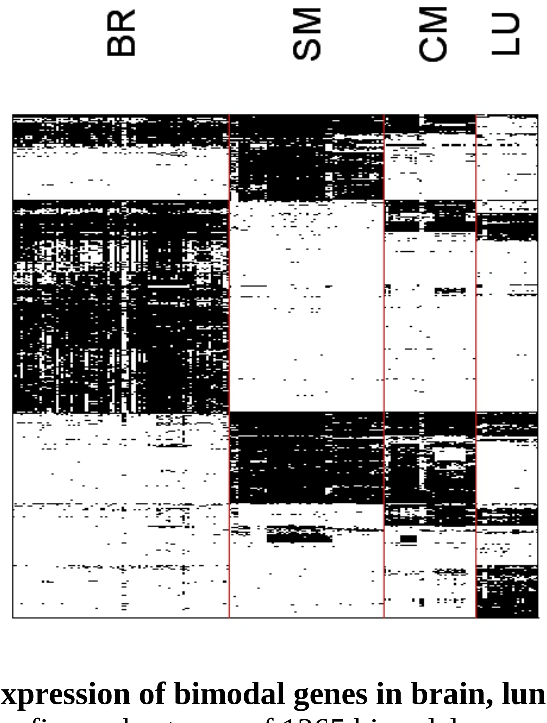


Figure 15 – Binarized expression of bimodal genes in brain, lung, skeletal muscle and cardiac muscle. Top figure: heat map of 1265 bimodal gene expression in 217 tissue samples. A black/white point at *i*,*j* indicates gene *i* is "on"/ "off" in sample *j*.

A black/white element x_{ij} of the heat map indicates gene *i* is expressed in the "on"/"off" mode in sample *j*. Distinct clusters of "on" and "off" genes are observed in each tissue type. In total, we identified 542, 429, 322, and 278 genes over-represented in the "on" mode and 645, 778, 830 and 896 genes over-represented in the "off" mode in brain, skeletal muscle, cardiac muscle and lung tissue respectively. Functional enrichment analysis with gene sets defined by GO terms and KEGG pathways (Tables 5 and 6) provided a biological context to these sets of bimodal genes. Notably, neural tissue-specific processes including neural migration, adhesion, recognition and differention, nervous system development, and synaptic transmission populate the list of GO terms associated with genes that are "on" in brain tissue. Similarly, terms related to muscle

Table 5 – GO categories significantly enriched with "on" genes in brain tissue
P-values <= 0.001 indicates significance.

Biological Process	Cellular Component	Molecular Function
▪ Neuron migration ▪ Transport ▪ Ion transport ▪ Negative regulation of microtubule depolymerization ▪ Cell adhesion ▪ Neuron adhesion ▪ Transmembrane receptor protein tyrosine phosphatase signaling pathway ▪ Synaptic transmission ▪ Neuromuscular synaptic transmission ▪ Nervous system development ▪ Synaptogenesis ▪ Central nervous system development ▪ Neuron recognition ▪ Anterograde axon cargo transport ▪ Neuron differentiation	▪ Cytoskeleton ▪ Microtubule ▪ Microtubule associated complex ▪ Neurofilament ▪ Membrane ▪ Integral to membrane ▪ Synaptosome ▪ Cell junction ▪ Axon ▪ Growth cone ▪ Synapse ▪ Postsynaptic membrane	▪ Actin binding ▪ GTPase activity ▪ Transmembrane receptor protein tyrosine ▪ Structural molecule activity ▪ Strucutural constituent of cytoskeleton ▪ Ion channel activity ▪ Structural constituent of myelin sheath

Biological Process	Cellular Component	Molecular Function
▪ Regulation of the force of heart contraction^{SM,CM} ▪ Glycolysis^{SM,CM} ▪ Tricarboxylic acid cycle^{SM,CM} ▪ Phosphate cycle^{SM,CM} ▪ Muscle contraction^{SM,CM} ▪ Striated muscle contraction^{SM,CM} ▪ Cytoskeleton organization and biogenesis^{SM,CM} ▪ Muscle development^{SM,CM} ▪ Regulation of heart contraction^{SM,CM} ▪ Muscle thin filament assembly^{SM,CM} ▪ Actomyosin structure organization and biogenesis^{SM,CM} ▪ Negative regulation of heart contraction^{SM,CM} ▪ Atrial cardiac muscle morphogenesis^{SM,CM} ▪ Carbohydrate metabolic process^{SM} ▪ Glycogen metabolic process^{SM} ▪ Glycogen biosynthetic process^{SM} ▪ Gluconeogenesis^{SM} ▪ Protein amino acid dephosphorylation^{SM} ▪ Regulation of muscle contraction^{SM} ▪ Regulation of striated muscle contraction^{SM} ▪ Somatic muscle development^{SM} ▪ Blood circulation^{SM} ▪ Dephosphorylation^{SM} ▪ Maintainance of epithelial cell polarity^{SM} ▪ Glycerol-3-phosphate catabolic process^{SM} ▪ Response to unfolded protein^{CM} ▪ Cell adhesion^{CM} ▪ Cell-matrix adhesion^{CM} ▪ Heart development^{CM} ▪ Adult heart development^{CM} ▪ ATP transport^{CM} ▪ Focal adhesion formation^{CM}	▪ Cytoplasm^{SM,CM} ▪ Smooth endoplasmic reticulum^{SM,CM} ▪ Cytoskeleton^{SM,CM} ▪ Striated muscle thick filament^{SM,CM} ▪ Actin cytoskeleton^{SM,CM} ▪ Sarcoglycan complex^{SM,CM} ▪ Sarcoplasmic reticulum^{SM,CM} ▪ Myofibril^{SM,CM} ▪ Sarcomere^{SM,CM} ▪ Z disc^{SM,CM} ▪ Mitochondrion^{SM} ▪ Mitochondrial inner membrane^{SM} ▪ Mitochondrial matrix^{SM} ▪ Muscle myosin complex^{SM} ▪ Troponin complex^{SM} ▪ Actin filament^{SM} ▪ Myosin complex^{SM} ▪ Sarcoplasmic reticulum membrane^{SM} ▪ Sarcoplasmic reticulum lumen^{SM} ▪ Proteinaceous extracellular matrix^{CM}	▪ Actin binding^{SM,CM} ▪ Citrate (Si)-synthase activity^{SM,CM} ▪ Electron-transferring-flavoprotein dehydrogenase activity^{SM,CM} ▪ Extracellular matrix structural constituent^{SM,CM} ▪ Calcium ion binding^{SM,CM} ▪ Structural constituent of muscle^{SM,CM} ▪ SSM00 alpha binding^{SM,CM} ▪ Microfilament motor activity^{SM} ▪ Motor activity^{SM} ▪ Catalytic activity^{SM} ▪ NADH dehydrogenase activity^{SM} ▪ Glycerol-3-phosphate dehydrogenase (NAD+) activity^{SM} ▪ Calmodulin binding^{SM} ▪ Tropomyosin binding^{SM} ▪ Electron carrier activity^{SM} ▪ Oxidoreductase activity, acting on CH-OH group of donors^{SM} ▪ Oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor^{SM} ▪ Hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances^{SM} ▪ Spectrin binding^{SM} ▪ NAD binding^{SM} ▪ Structural molecule activity^{CM} ▪ Structural constituent of cytoskeleton^{CM} ▪ Protein binding^{CM} ▪ Adenine transmembrane transporter activity^{CM}

Table 6 – GO categories significantly enriched with “on” genes in skeletal and/or cardiac muscle

P-values <= 0.001 indicate significance. ^{SM}skeletal muscle, ^{CM}cardiac muscle

development and organization, muscle contraction, calcium ion binding, cellular metabolism and muscle-specific structures such as the sarcoplasmic reticulum, myofibril, sarcomere and z disc are found in the list of enriched GO terms associated with skeletal and cardiac muscle. A number of KEGG pathways are also enriched (Table 7). The KEGG diagram summarizing cell adhesion molecules is enriched with genes turned "on" in brain tissue and genes turned "off" in muscle tissue (Figure 16). Several of these cell adhesion molecules, such as CDH2, NCAM, NRXN, and NLGN, are expressed at

Table 5 – KEGG pathways enriched with “on” genes in brain, skeletal muscle, cardiac muscle and lung tissue

P-values <= 0.001 indicate significance. ^{B}brain, ^{SM}skeletal muscle, ^{CM}cardiac muscle, ^{L}lung

KEGG Pathways
▪ Cell adhesion molecules (CAMs)^{B}
▪ Long-term depression^{B}
▪ Neurodegenerative diseases^{B}
▪ Tight junction^{B,SM}
▪ Calcium signaling pathway^{SM}
▪ Carbon fixation^{SM}
▪ Citrate cycle (TCA cycle)^{SM}
▪ ECM-receptor interaction^{SM,CM,L}
▪ Focal adhesion^{SM,CM,L}
▪ Glycolysis / Gluconeogenesis^{SM,CM}
▪ PPAR signaling pathway^{SM,CM}
▪ Reductive carboxylate cycle (CO2 fixation)^{SM,CM}
▪ Cell Communication^{CM,L}
▪ Pyruvate metabolism^{CM}
▪ Adherens junction^{L}
▪ Complement and coagulation cascades^{L}

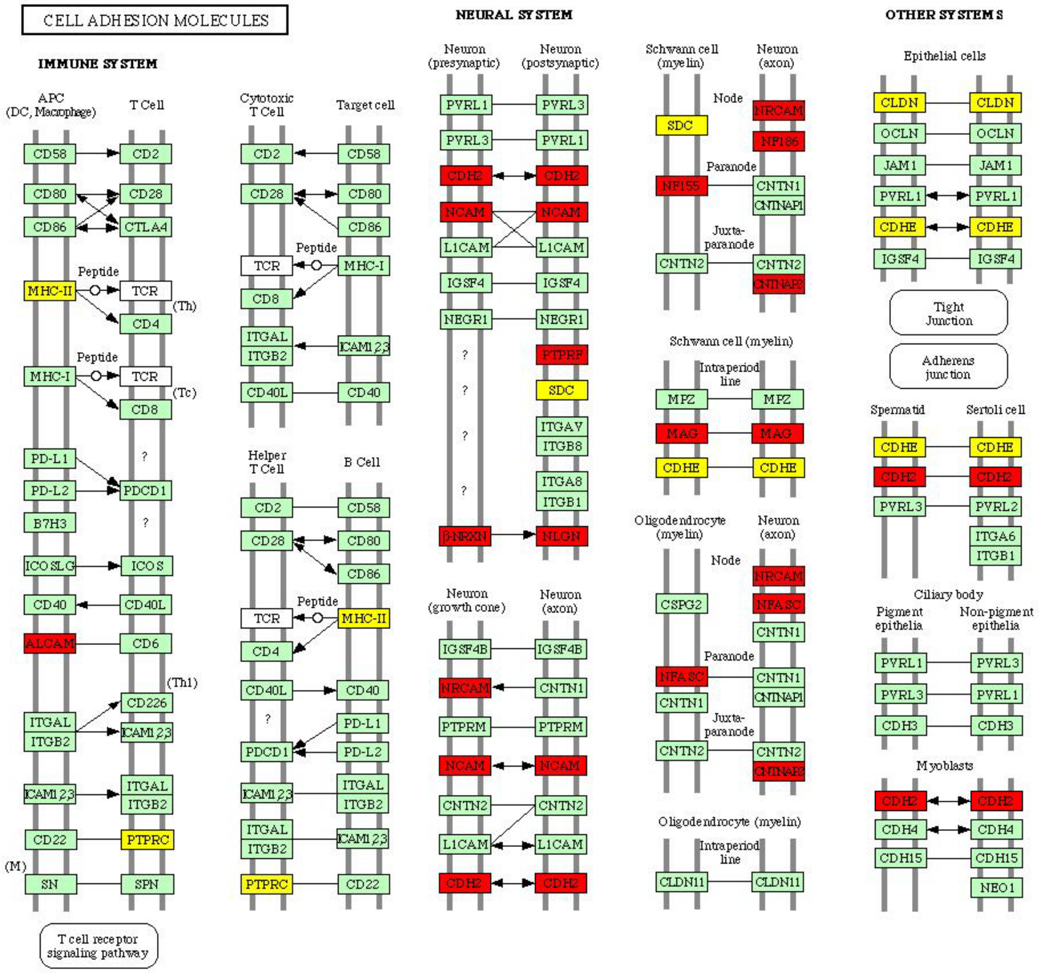


Figure 16: Bimodal gene expression in KEGG cell adhesion molecules diagram. Genes marked with red are "on" in brain tissue and "off" in muscle tissue. Genes marked with yellow are "off" in muscle tissue.

synaptic junctions [157]. Another subset, including NFASC and CNTNAP2, is integral to the formation of myelinated neurons [158]. Statistical enrichment of GO terms and KEGG pathways associated with tissue-specific structure and function provides further evidence that our bimodal gene set exhibits tissue-specific expression patterns.

5.4.5 Model-based classification of infectious disease and immune response signature

Model-based clustering of bimodal gene expression led to accurate classification of disease phenotypes in 221 microarray tissue samples profiling infectious disease.Peripheral blood mononuclear cells (PBMC) present in the circulation and lymphatic system recognize pathogen-specific molecules and initiate the immune response [159]. Pathogen recognition induces transcriptional activation of several host defense signaling pathways [160]. The posterior pairwise probability matrix derived from model-based clustering partitioned expression profiles of PBMCs into disease-specific clusters for HIV-1 infection, hepatitis C, influenza, and malaria (Figure 17). Moreover, model-based clustering differentiated between samples of hepatitis C infection in PBMCs and liver biopsies (Figure 17). These results suggest that model-based clustering captures infectious disease signatures in microarray data in a tissue-specific manner. In addition, a different set of bimodal genes may be selectively expressed in PBMCs in infectious disease states induced by different pathogens.

Enriched functional gene sets related to the immune response were detected in sets of active switch genes in infectious disease samples. Of the 1295 bimodal genes analyzed, 192, 160, 148 and 117 genes were expressed in the “on” mode in the majority of samples from PBMCs in hepatitis C, influenza A, malaria, and HIV respectively. In liver biopsies from hepatitis C infected individuals, 301 bimodal genes are over-represented in the “on” mode. Table 6 lists the GO terms that are statistically enriched in Hepatitis C, influenza, and malaria infection. Biological processes commonly enriched in the set of bimodal genes expressed in the “on” mode in these diseases include B cell receptor

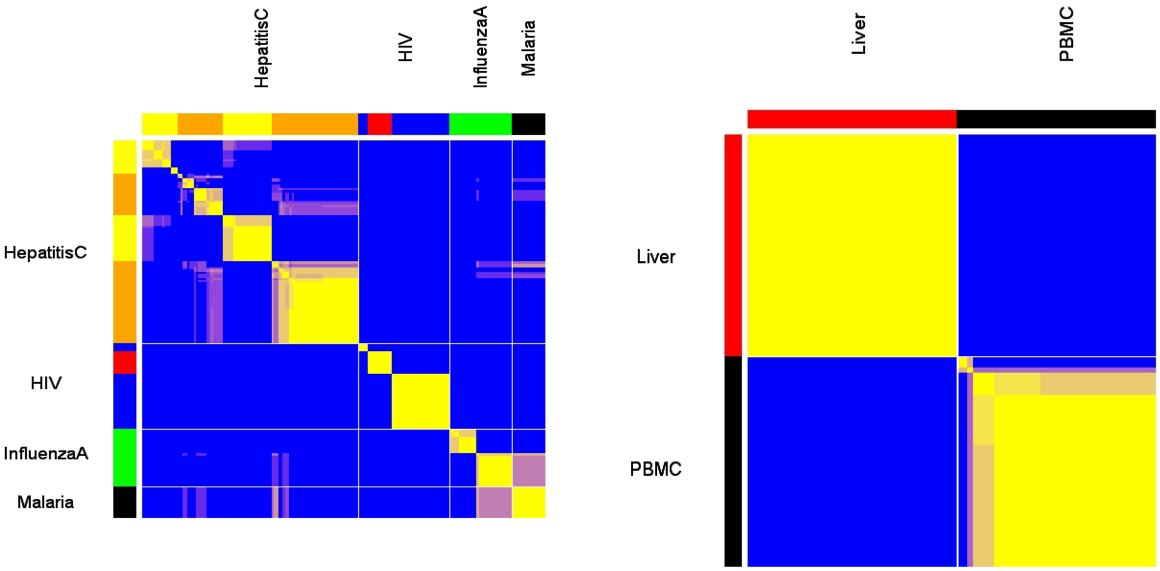


Figure 17 – Model-based clustering of bimodal gene expression classifies infectious disease states separately and identifies tissue-specificity in hepatitis C infection. Heat map representation of pairwise posterior probabilities derived from model-based clustering of infectious disease expression data. Left column: Classification of hepatitis C, HIV, influenza A, and malaria profiled in peripheral blood mononuclear cells (PBMCs). Right column: Classification of hepatitis C infection profiled in peripheral blood mononuclear cells and liver biopsies.

signaling [161, 162] and humoral immune response involving circulating immunoglobulins [163]. These processes are central in the activation of the antigen-mediated, adaptive immune system. Bimodal genes upregulated in hepatitis C infection in PBMCs are associated with inflammatory response, respiratory burst and altered response to calcium ions (Table 8) [164]. Both inflammation and the production and release of oxidative species are important components of the innate immune response [165]. Enrichment of gene sets associated with function of the immune system and leukocyte-specific receptor signaling pathways suggests that a subset of genes with bimodal expression patterns are relevant in the host-response to pathogens.

Table 8 – GO categories significantly enriched with “on” genes in infectious disease
P-values <= 0.001 indicate significance in malaria, influenza A, hepatitis C-PBMCs and hepatitis C-Liver. P-values <= 0.01 indicate significance in HIV. [1]malaria, [2]influenza A, [3]HIV, [4]hepatitis C-PBMC, [5]hepatitis C-liver

Biological Process	Cellular Component	Molecular Function
▪ Immune response[1,2,3,4,5] ▪ Humoral immune response by circulating immunoglobin[1,2,4,5] ▪ Positive regulation of B cell proliferation[1,2,4,5] ▪ Early endosome to late endosome transport[1,2,4,5] ▪ Positive regulation of peptidyl-tyrosine phosphorylation[1,2,4,5] ▪ B cell receptor signaling pathway[1,2,4,5] ▪ Activation of MAPK activity[1,2,4] ▪ tRNA aminoacylation for protein translation[1,4] ▪ Antigen processing and presentation[1,4] ▪ DNA methylation[3] ▪ Translational initiation[3] ▪ Negative regulation of protein kinase activity[3] ▪ Defense response[3] ▪ Inflammatory response[4] ▪ Hemocyte development[4] ▪ Cell-cell adhesion[4] ▪ Pyridine nucleotide biosynthetic process[4] ▪ Respiratory burst[4] ▪ Response to calcium ion[3,4] ▪ Tricarboxylic acid cycle[5] ▪ Cell adhesion[5] ▪ Blood coagulation[5] ▪ Sensory perception of sound[3,5]	▪ B cell receptor complex[1,2,4,5] ▪ Immunoglobulin complex, circulating[1,2,4,5] ▪ Perinuclear region of cytoplasm[1,2,4,5] ▪ External side of plasma membrane[1,4] ▪ Membrane fraction[4,5] ▪ Cytoplasm[3,5] ▪ Cytoskeleton[3] ▪ Actin cytoskeleton[3] ▪ Extracellular region[5] ▪ Proteinaceous extracellular matrix[5] ▪ Collagen[5]	▪ Antigen binding[1,2,4,5] ▪ Succinate dehydrogenase activity[2,3,4] ▪ RNA binding[3] ▪ Structural constituent of cytoskeleton[3] ▪ Protein binding[3] ▪ Electron-transferring-flavoprotein dehydrogenase activity[5] ▪ Endopeptidase inhibitor activity[5] ▪ Structural molecule activity[5] ▪ Extracellular matrix structural constituent[5]

Gene Ontology enrichment analysis for switch-like genes turned “on” in HIV1 infection indicated the biological processes of DNA methylation, translational initiation, negative regulation of protein kinase activity, and response to calcium (Table 8). Statistically enriched KEGG pathways for HIV-1 infection included focal adhesion and adherens junction, leukocyte migration, natural killer cell mediated cytotoxity, B-cell and T-cell receptor signaling pathways. Bimodal genes that are observed to be expressed in the “on” mode in the T-cell signaling pathway include membrane protein CD45 [166], kinase activator SLP-76 [167], RAS proteins RASGRP1 and Rho Cdc42, calcium are involved in binding protein CaN, and the transcription factor AP1 [168] (Figure 18). These proteins multiple pathways and processes including ubiquitin mediated proteolysis, regulation of actin cytoskeleton, and proliferation and differentiation of the immune response. Many of these processes/genes are involved in the hijacking of normal T-cell function by HIV for the production, modification and release of viral proteins.

5.4.6 Supervised classification with bimodal genes capture tissue-specificand infectious disease specific signatures in microarray data

We implemented a multi-class supervised classification scheme to estimate whether tissue/infectious disease-specific bimodal gene expression signatures were conserved in independent data. Each dataset was split into training and test sets in a class-proportional manner. Training data was used to select the 5 most discriminative switch-like genes and generate multiple binary decision rules. Each decision rule was trained to recognize one class versus all others [148]. Test samples were classified with the decision rules trained on independent data to provide an unbiased evaluation of the association of bimodal gene expression with class distinction. As a control, we also trained classifiers on the basis of genes selected randomly from the entire gene chip. Results over 100 independent iterations of training and testing are shown in Tables 9 and 10. Prediction of tissue-

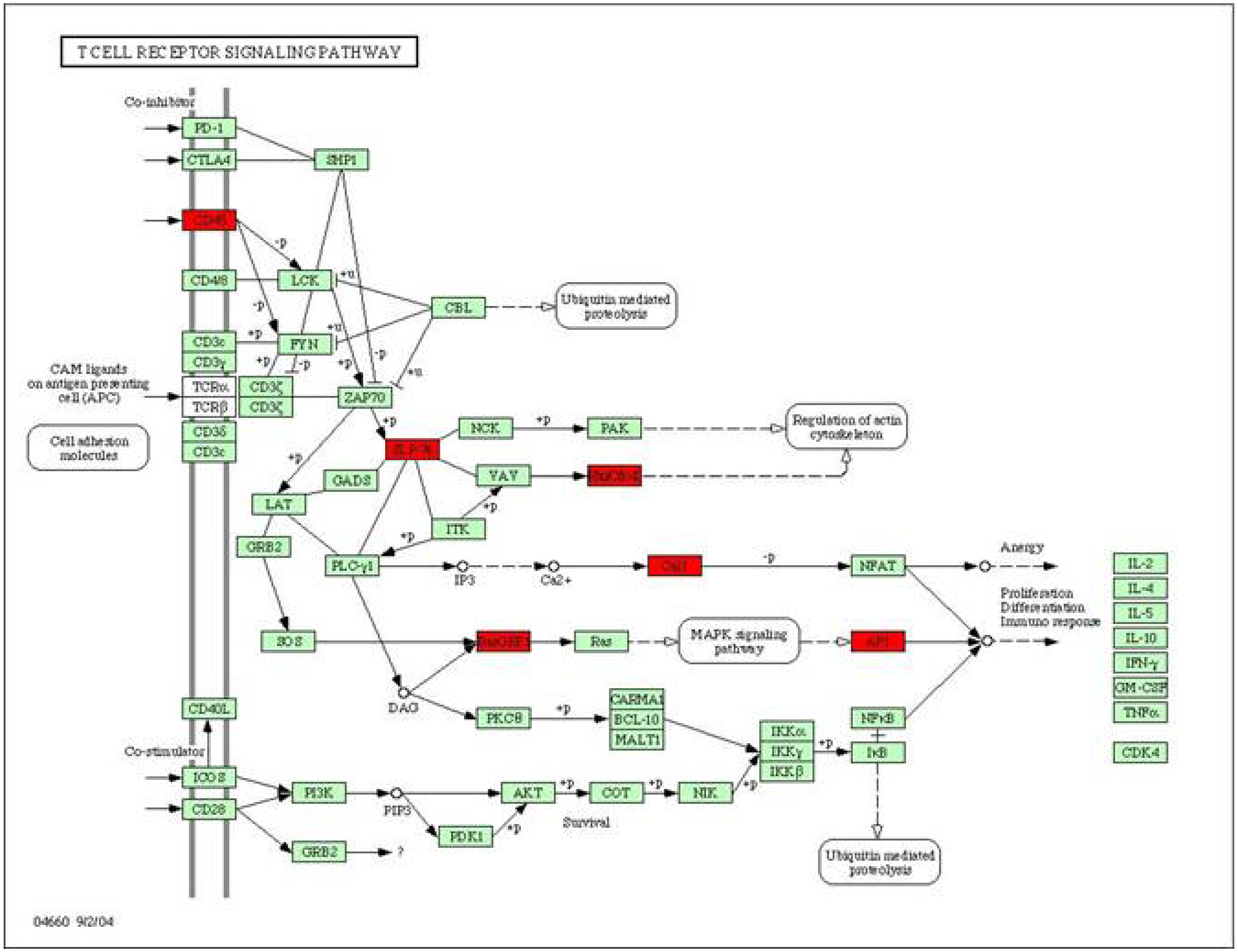


Figure 18 – Bimodal genes that were switched "on" as a result of HIV infection in KEGG T-cell receptor signalling pathways. Bimodal genes marked with red are "on" in the KEGG T-cell receptor signaling pathway in HIV infection.

specificity was accurate in 85 % of test samples for all tissues except colon (10 samples), mammary (15 samples), small intestine (7 samples) and testis (38 samples). Misclassified tissue samples were often classified as similar tissue types. For example, microarray samples from small intestine tissue were predicted to be either muscle tissue or pancreatic tissue in 30% and 24% of test samples respectively. These results indicate the persistence of cell-type-specific expression signatures in heterogeneous tissue samples. Notably, 14% of testis samples were misclassified as ovary, indicating a subset of bimodal genes may be similarly expressed in reproductive organs of the male and female. Supervised classification of infectious diseases based on switch-like genes showed similar accuracy (Table 10). Multi-class supervised classification separated microarray samples from HIV-1 infection, hepatitis C and malaria well but it has allocated 22% of the influenza microarray samples to the bin for hepatitis C. This is not surprising in the light of our findings showing common immune signaling responses for these two viral infections (Table 8). In both the classification of tissue phenotypes and infectious disease, feature selection was more accurate than random selection. These results indicate that tissue-specific and disease-specific bimodal gene expression profile signatures are conserved in independent data.

5.4.7 Effect of sample size, effect size and number of informative genes on classification accuracy

Supervised classification of simulated gene expression profiles illustrated the strong dependence of prediction accuracy on sample size, effect size and the number of informative genes (Figure 19). In this minimal model, simulated datasets were designed to approximate binary classification. A response variable indicative of class membership

Table 9 – Classification accuracy in supervised clustering of tissue phenotypes

Values equal the proportion of true class versus predicted class membership over 100 iterations of training and testing. Values representing correct classification are outlined in bold.

		Predicted Class																		
		adipose	adrenal	brain	colon	epidermal	heart	kidney	liver	lung	mammary	muscle	ovary	pancreas	peripheral_blood	small_intestine	spleen	stomach	testis	thymus
True Class	adipose	0.89				0.11														
	adrenal		1																	
	brain		0.01	0.99																
	colon				0.77	0.12								0.02		0.06			0.01	
	epidermal	0.01				0.97					0.02									
	heart						1													
	kidney	0.08						0.86				0.01	0.01	0.04						
	liver								1											
	lung									0.97						0.01		0.01		0.01
	mammary	0.08				0.13					0.79									
	muscle											1								
	ovary	0.02											0.97						0.01	
	pancreas													0.91		0.06		0.02		
	peripheral_blood														0.88		0.09			0.03
	small_intestine											0.3		0.24		0.44	0.01	0.01		
	spleen												0.01		0.04	0.02	0.92			
	stomach				0.01		0.03									0.02		0.95		
	testis				0.05						0.01		0.14			0.01		0.03	0.73	0.03
	thymus															0.12	0.01			0.86

Table 10 – Classification accuracy in supervised clustering of infectious disease

Values equal the proportion of true class versus predicted class membership over 100 iterations of training and testing.

		Predicted Class				
		HepatitisC_PBMC	HepatitisC_Liver	HIV	InfluenzaA	Malaria
True Class	HepatitisC_PBMC	0.91			0.02	0.06
	HepatitisC_Liver		1			
	HIV			1		
	InfluenzaA	0.22			0.72	0.06
	Malaria	0.01			0.02	0.97

was generated from expression profiles of informative genes via logistic regression. Informative gene expression profiles were drawn from a pair of multivariate normal distributions. Expression profiles representing noise genes were generated from a mixture of these distributions. Separation between the two distributions was defined in terms of the difference in class-specific means and the pooled variance. Bimodal gene expression was assumed to hold when setting the separation equal to the median separation between “on” and “off” modes of switch-like expression profiles (μ_1-μ_2 = $6\sigma^2$). Smaller separation values (μ_1-μ_2 = $2\sigma^2$, μ_1-μ_2 = σ^2) simulate expression profiles that are less bimodal and more normally distributed (Figure 19). Supervised classification was applied as described above and classification accuracy was assessed using the area under the receiver-operating characteristic curve (AUC) (Figure 19). Results are based on the average AUC generated from 100 simulated datasets for each condition. Classification accuracy generally improved as expression profiles became more bimodal (Figure 19). Increased sample size and decreased number of informative genes also resulted in more accurate classification as well (Figure 19).

5.5 Discussion

Development and subsequent commercialization of microarray platforms has led to extensive investigation of global gene expression profiles in health and disease. Expression profiling of diverse healthy tissues provides a comprehensive perspective of the range of transcriptional regulation under physiologic conditions [169-171]. Similarly, identification of gene expression signatures indicative of disease subtypes improves our understanding of the molecular basis of pathology [135, 136]. Small sample size and the

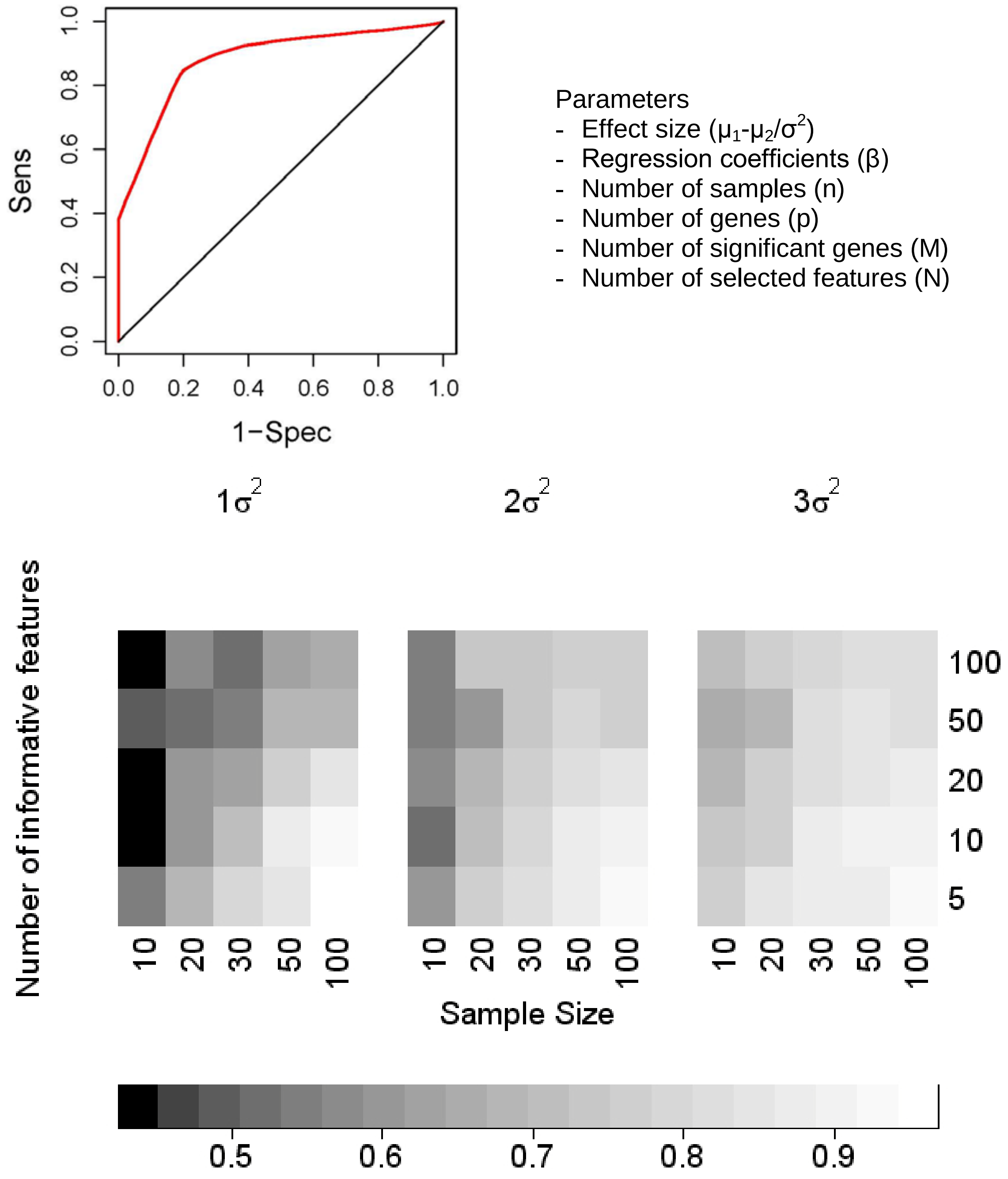


Figure 19 – Effect of sample size, separation and number of informative genes on classification of simulated expression data. Classification accuracy is measured with the area under the receiver operating characteristic curve.

large number of measurements for each sample are among the limiting factors that hinder the effectiveness of gene expression profiling and drive the development of new analytical methods.

Unsupervised clustering of microarray data classifies samples in an unbiased manner according to similarity in gene expression profiles. Adaptation of model-based clustering to low sample size, high dimensional datasets [66] and formalization of statistical approaches for selecting the optimum number of clusters [146] represent significant advances. In this study, we used these advanced methods to cluster and classify infectious disease and tissue phenotypes in large scale microarray data using a reduced set of 1265 genes, the so-called switch-like genes [134]. Switch-like genes are identified through the detection of bimodal gene expression patterns across diverse biological conditions. Switch-like genes are likely to be under strict transcriptional regulation and are statistically enriched for cell membrane and extracellular proteins [144].

We demonstrated that model-based clustering of switch-like gene expression patterns differentiates between tissue phenotypes in a microarray dataset with tissue-specific sample sizes ranging from 5 to nearly 100. Model-based clustering operates on the assumption that samples are drawn from multivariate Gaussian distributions. Clusters are defined by identifying shifts in the mean expression value for a subset of genes. Based on this description, model-based clustering is particularly well-suited for the analysis of bimodal gene expression profiles. Annotation of genes with cellular localization information allowed us to identify a subset of 300 bimodal genes expressed on the extracellular matrix or the plasma membrane. This set includes membrane-bound integrin proteins and ECM proteins belonging to collagen, laminin, and fibronectin families. Accurate classification of tissue type with this subset of bimodal genes supports the hypothesis that interaction with the cellular micro-environment has a significant role in tissue differentiation [151, 152]. Distance-based unsupervised classification methods such as Kmeans and hierarchical clustering also identified brain-specific and muscle-specific clusters with sample sizes above 40 but they tended to group tissues with few microarray samples together. Classification accuracy will likely improve with increasing sample size.

Model-based clustering of the set of 1295 bimodal genes correctly placed microarray samples into bins identified for HIV-1 infection, hepatitis C, influenza and malaria. In this classification, each disease type might have multiple bins depending on tissue type and/or laboratory from which microarray data came from. The method classifies hepatitis C microarray samples from liver biopsies into a separate bin rather than mixing it with microarray data on peripheral blood cells from hepatitis C patients. Similarly, microarray data on HIV-1 infection turned out to be classified into laboratory-specific bins. This differentiation may be due to distinctly different patient pools in different laboratories as well as the small sample size in disease microarray sets. Nonetheless, these results indicate the promise of model-based classification in the identification of infectious disease subtypes from microarray data.

Identification of on-off states of switch-like genes in microarray data allowed us to assess the biological relevance of the alternate switch states of these genes in various infectious diseases and tissue phenotypes. Comparison of activated switch-like gene sets between tissue and disease phenotypes provide a measure of distance between different phenotypes. We observed that genes expressed in the "on" mode in brain tissue and the "off" mode in muscle tissue code for neural-specific cell adhesion molecules. In addition, bimodal genes switched "on" in brain, skeletal muscle and cardiac tissue are related to tissue-specific structure and function. In the infectious disease states investigated here, bimodal genes expressed in the "on" mode are related to both innate and antigen-mediated immune responses. Additionally, in HIV samples, "on" genes are expressed in pathways related to the hijacking of infected T-cells for viral production. The large body of evidence presented in the results section points to the success of switch-like gene sets in capturing biologically-relevant global gene expression signatures from microarray data.

Given the demonstrated biological relevance of bimodal expression patterns, it would be worthwhile to determine the clinical relevance of switch-like gene annotation. Identification of bimodal genes expressed in the on state in complex diseases such as autism, diabetes and cancer may provide a method for dimension reduction in the identification of disease-related single nucleotide polymorphisms (SNPs) [34, 172-175] and expression quantitative trait loci (eQTL) [176, 177] in genome-wide association studies. Both gene sequences and promoter regions of on switch genes as determined from large scale microarray data could be searched for SNPs and eQTL linked to the onset of disease or disease progression. Further studies are needed to investigate the full potential of clinically relevant classification using switch-like gene annotation from microarray data.

We also addressed the question of whether tissue-specific or infectious disease bimodal expression signatures are conserved in independent data. Unsupervised clustering algorithms use all of the data to identify similar expression profiles: these algorithms may reveal patterns associated with random noise. Supervised classification algorithms test for random associations by separating the data into independent training and testing sets. A multi-class supervised classification scheme was implemented. In both tissue phenotype and infectious disease datasets, a majority of test samples were correctly classified using as few as five genes. Classification on the basis of discriminative bimodal genes was more accurate than classification by control sets of genes selected randomly from the entire microarray chip. Moreover, our simulation results presented in Figure 19 indicate that tissue and infectious disease specific bimodal expression signatures are likely to be conserved in independent data at large sample sizes.

5.6 Conclusion

In this study, we used advanced clustering and classification algorithms to investigate expression profiles of switch-like genes in multiple tissue and infectious disease phenotypes. Switch-like genes are defined as those genes with bimodal expression patterns in large-scale microarray data containing hundreds of samples across different tissue types. Use of a model-based clustering algorithm accurately classified more than

400 microarray samples into 19 different tissue types on the basis of bimodal gene expression. The algorithm demonstrated similar accuracy in the classification of microarray data corresponding to hepatitis C, influenza, HIV-1 infection and malaria. Classification accuracy was exceptional even with class-specific sample sizes between ten and twenty arrays. Supervised classification with feature selection restricted to switch-like genes also recognized tissue-specific and infectious disease specific signatures in independent test datasets reserved for validation. Moreover, our computational simulations with a minimal model of microarray data indicated the validity of our observations in a large number of circumstances. A set of 300 genes out of the 1295 genes annotated in the human as switch-like coded for either extracellular matrix or cell membrane proteins. This subset was equally good in differentiating distinct tissue types, indicating a potential role for them as biomarkers provided that expression is altered in the onset of disease. Determination of “on” and “off” states of switch-like genes in various tissues and diseases allowed for prediction of activated/deactivated genes/pathways that are consistent with existing research data. Future work is needed to address the question of whether switch-like gene expression has clinical implications in disease subtype classification.

Chapter 6: Identification of autism risk loci around neural-specific bimodal genes

6.1 Summary

This chapter discusses an association study of high-density genomic data obtained from a multiplex cohort of 189 families affected by autism compiled by the Autism Genetic Resource Exchange (AGRE). Autism is a heterogeneous neurodevelopmental disorder that is characterized by impaired social interaction and communication and repetitive behavioral patterns. Epidemiological evidence suggests a strong heritable component transmitted by multiple genetic loci. Candidate gene regions likely to contain genetic variants associated with autism risk were identified using gene expression analysis of a microarray dataset profiling 19 different tissue types. We defined a set of genes with bimodal expression patterns across all tissues and high levels of expression in a majority of brain samples as neural-specific switch-like genes. The coding and cis-regulatory regions of these genes were used as candidate gene regions. Cis-regulatory regions were conservatively identified using a 1 Megabase window centered at the midpoint of the gene coding region. Autistic individuals in this study were identified by positive diagnosis from the Autistic Diagnostic Interview-Revised (ADIR) and the Autism Diagnostic Observation Schedule (ADOS). A two-stage family-based association test (FBAT) strategy was used to test for association and correct for multiple testing. With this procedure, we identified a single nucleotide polymorphism (refSNP identifier: rs17101921) associated with autism with genome-wide significance in the q26 region of chromosome 10. Subjects with the A allele at this locus are more likely to be diagnosed with autism (odds ratio = 1.31, 95% confidence interval (0.81 – 2.11). Although none of the other screened SNPs in the region demonstrated association with autism, linkage

disequilibrium analysis identified a 13 Kilobase haplotype block containing the rs17101921 SNP. The rs17101921 single nucleotide polymorphism (SNP) is located approximately 80 Kilobases upstream of the fibroblast growth factor 2 gene (FGFR2). FGFR2 is highly expressed in glial cells of the central nervous system and is involved in nervous system development and repair after injury. Our study presents a novel method for integrating information obtained from gene expression and genotype analysis. Results of our study suggest new experiments regarding the investigation of the molecular basis of autism.

6.2 Background

Autism is one of a spectrum of neurological disorders that present with a combination of impaired social interaction, difficulties with communication, and repetitive behavior patterns. Autism has been linked to several environmental and genetic risk factors. . Approximately 10-15% of autism cases can be linked to chromosomal abnormalities such as fragile X syndrome, tuberous sclerosis or rare single gene disorders [178]. Epidemiological studies of disease concordance in familial and twin studies indicate a heritable genetic component of disease. Prevalence in males is four times higher than in females [179], suggesting that autism risk may be partially transmitted by loci on the X chromosome. Sibling recurrence risk (5-10%) is significantly greater than prevalence in the general population (0.15-0.2%) [180]. In addition, identical twins show much higher concordance (60%-92%) than fraternal twins (0-10%) [181]. Taken together, these findings suggest that autism susceptibility is partially conferred by variation at multiple genetic loci.

Several independent genetic linkage and association studies have been conducted to identify genetic variants associated with autism susceptibility. In linkage analysis, a set of genetic markers spaced widely throughout the genome are sequenced to detect regions of co-transmitted loci at low resolution. Studies of extensive pedigrees of autism affected families have identified linkage regions in the short arms of chromosomes 2, 3, 6, 7, 10 and 17 [182-185]. Association studies identify alleles at specific loci that are observed in affected individuals more than expected by chance. Fine-mapping of candidate genes in chromosomal regions identified from linkage analysis has identified a number of putative autism susceptibility loci. Genetic variants in the transcript region of the glutamate receptor 6 (GluR6) gene have been associated with increased autism risk [186]. The glutamate receptor functions in the excitation of neural signaling at post-synaptic junctions [187]. Similarly, variants in the laminin beta-1 (LAMB1) and engrailed 2 (EN2) genes, have been associated with autism. Products of both of these genes participate in the regulation of neurodevelopment [188, 189]. Additionally, autism-associated polymorphisms have been detected in the upstream regulatory and intronic regions of the serotonin-transporter gene (SLC6A4). SLC6A4 regulates the effect of serotonin by re-absorbing the neurotransmitter from the synaptic cleft [187]. Significant associations have been reported at several more genetic loci but these results have not been replicated in follow-up studies.

Recent technological developments and the identification of common genetic variants, or single nucleotide polymorphisms (SNPs), make it possible to survey genetic variation at relatively high resolution. Single nucleotide polymorphisms (SNPs) are individual nucleotide bases in the genetic code that vary from person to person. Extensive genetic mapping has identified a set of common SNPs expressed in at least 5% of the population [23, 24]. With this knowledge, gene mapping microarray platforms capable of evaluating association at a genome-wide scale have been developed. SNP profiling technologies such as these can be used to provide a more detailed evaluation of the common genetic variants involved with autism risk.

Genome-wide association studies are burdened with an exceedingly large multiple testing problem. Procedures which control the type I error rate are applied to adjust p-values for the number of hypotheses tested and maintain genome-wide significance. Conservative multiple testing corrections such as the Bonferonni correction are likely to eliminate a high number of true positive associations [56, 57]. The severity of multiple testing corrections can be reduced by decreasing the number of SNPs prior to association testing. A priori biological knowledge can be used to identify genes that are suspected to be involved with disease processes. Association tests are then limited to the coding and regulatory regions of these candidate genes. In family-based studies, statistical methods have been developed to screen for SNPs on the basis of conditional power estimates [190]. Only the most promising SNPs are tested for association. By reducing the number of hypotheses tested, genome-wide significance can be assessed at less stringent thresholds.

In this study, we have detected putative autism susceptibility loci in the coding and cis-regulatory regions of candidate genes identified from gene expression analysis. In previous work, we identified 1265 genes with bimodal or switch-like expression patterns in diverse human tissues. A subset of these bimodal genes are over-expressed in brain tissue and are known to be involved with neural development and function. Using genotype and phenotype data of individuals in 189 autism-affected families compiled by the Autism Genetic Resource Exchange (AGRE) [191], we scanned the coding and regulatory regions of these genes for genetic loci associated with autism susceptibility. A two-stage family-based association test was applied to screen for promising SNPs on the basis of conditional power estimates and test for association. Our scan of the candidate gene regions identified a SNP upstream of the fibroblast growth factor receptor 2 (FGFR2) gene associated with autism risk. FGFR2 is known to have important roles in neurodevelopment [192] but has not previously been associated with autism through genomic study. The autism susceptibility locus identified in this study provides evidence supporting novel hypotheses regarding the molecular origins of autism.

6.3 Methods and materials

6.3.1 Gene expression, genotype and phenotype data

Gene expression data of approximately 400 samples from 19 different tissue phenotypes was compiled (Table 2). Expression profiles were generated using the HGU133A and HGU133Plus2 Affymetrix platforms. Only probesets common to both arrays were retained, leaving 22277 probesets for downstream analysis.

Genotypic and phenotypic data compiled from family-based studies of autism were obtained from the Autism Genetic Resource Exchange (AGRE) [191]. Approval was obtained from the institutional review board of Drexel University prior to requesting the data. Genotypes were generated using the Affymetrix Genome-Wide Human SNP Array 5.0 platform which contains probesets representing approximately 400,000 SNPs. The subject population consists of 721 simplex and multiplex pedigrees with approximately 1385 affected individuals. Subjects were screened for common chromosomal aberrations associated with autism, such as Fragile X syndrome. Subjects with chromosomal abnormalities or other non-idiopathic conditions were excluded to reduce phenotypic heterogeneity. Additionally, families were excluded on the basis of incomplete genotyping. Only families in which both parents and one or more affected children were genotyped were included. To account for population stratification, only self-identified Caucasian subjects were included. Filtering by chromosomal abnormalities, incomplete pedigree, and race resulted in the exclusion of 250 families.

Diagnosis of autism is based largely on the Autism Diagnostic Interview-Revised (ADI-R) algorithm [193]. In addition, the Autism Diagnostic Observation Schedule (ADOS) can be used to distinguish between autism and other pervasive developmental disorders such as Asperger’s and PDD-NOS [194]. In our analysis, affected individuals were identified by autism diagnoses by both the ADI-R and ADOS. Furthermore, only families with two or more affected individuals were included. Use of the multiplex diagnostic specifications described above resulted in the inclusion of 189 affected families, with 808 subjects and 392 autistic subjects.

6.3.2 Pre-processing and quality control

Reference robust multi-chip averaging (refRMA) [50] was used for normalization of microarray data, as described in the analysis of switch-like expression patterns. Briefly, RMA background adjustment was applied to each array, arrays are normalized with quantile normalization on the basis of a reference empirical distribution derived from a biologically diverse training set of arrays, and normalized probe intensities are adjusted to account for probe affinity effects derived from the training data. Following background correction and normalization, summarized probe set expression values are obtained from the median value of constituent probe level intensities.

Raw data derived from gene mapping arrays was pre-processed by the AGRE using the Birdseed algorithm. Birdseed is used for normalization and summarization of probe-level data and genotype calling [51, 195]. Briefly, quantile normalization is used to correct for chip-specific effects. A log transformation is used to obtain corrected log-scale probe-level intensity measures. Median polish is used to adjust for probe-specific effects and derive summarized allele-specific signal values (A and B alleles). A model-based clustering approach is used to estimate the genotype of each sample at each SNP location. Clusters of genetic loci are generated by plotting the fluorescent signal derived from A allele probes versus the signal from B allele probes. Gentoypes are called by fitting SNP-specific Gaussian mixture models to the signal values in this two-dimensional space using expectation maximization. Following normalization and genotype calling, a number of quality control measures were implemented with the PLINK program [196] to screen for potential genotyping errors. Mendelian inconsistencies in family pedigrees were identified and eliminated by setting the corresponding alleles to missing. Exact tests of Hardy-Weinberg equilibrium (HWE) were evaluated for each SNP based on the genotype of the founders in each pedigree. Loci with > 10% missing values, HWE p-values < 0.001 or minor allele frequencies < 0.05% were discarded prior to linkage analysis. Implementation of quality control measures resulted in the removal of approximately 15% of the genotyped SNPs.

6.3.3 Identification of candidate autism susceptibility loci

Candidate genes with a priori relevance in neural development and function were identified using a statistical method to detect bimodality [140, 141, 144] in gene expression patterns in diverse human tissue. Bimodal genes were identified using a log-likelihood ratio test to test the alternative hypothesis that expression distributions fit a two-component Gaussian mixture model (GMM) versus a null normal distribution. P-values were obtained from evaluating the chi-square distribution at the values of the test statistic with six degrees of freedom. Genes with p-values < 0.001 and a standardized area of intersection between the distributions of the component Gaussians less than or equal to 0.01 were considered bimodal [144]. Genes with brain-specific expression patterns were identified by binarizing expression values with thresholds defined at the intersection of the probability density functions of the GMM for each gene [144]. Expression values above this threshold are described as "high" or "on". For each gene, each observation or sample was modeled as an independent trial in which success was defined as expression in the "on" mode. P-values were calculated from the binomial distribution with an equal probability of success and failure. P-values less than or equal to 0.01 indicates a significant association between bimodal gene expression and phenotype. Approximately 542 genes were expressed in the “on” mode in a majority of neural tissue samples.

Single nucleotide polymorphisms in the coding and regulatory regions of candidate genes were identified using annotation information obtained from the NetAffx database maintained by Affymetrix. Probesets on the HGU133A and HGU133plus2 platforms mapping to multiple chromosomal regions were excluded from analysis. This filtering step resulted in the exclusion of 35 genes. Chromosomal regions containing coding and cis regulatory regions of candidate genes were conservatively identified using a 1 Mb window centered at the midpoint of the sequence at which each probeset aligns [177]. In this manner, 55,214 SNPs in candidate gene regions were identified.

6.3.4 Two-stage family-based association test

Family-based association tests (FBATs) use genotypes of parents and affected offspring to evaluate the composite null hypothesis of no linkage between disease loci and tested loci and no association between genotype and disease phenotype. The test statistic is a generalization of the transmission disequilibrium test (TDT) that compares the frequency at which alleles are passed to affected offspring with its expected value derived from parental genotypes [197]. Only families with heterozygous parental genotypes are informative in the calculation of the test statistic [198]. Given a genetic locus with two alternate alleles A and B, the FBAT statistic is calculated as follows:

$$\begin{aligned} U &= (S - E[S])/\sqrt{V} \\ S &= \sum_{ij} T_{ij} X_{ij} \\ V &= Var(S) \end{aligned}$$

where T_{ij} represents the phenotype (ie. T_{ij} = 1 if affected, 0 otherwise) and X_{ij} represents the genotype of the *i*th offspring in the *j*th family [190]. The value of X_{ij} is dependent on the genetic model being evaluated. For example, under an additive model X_{ij} is equal to the number of A alleles (i.e. 0, 1, or 2) in the genotype of the *ij*th individual [74]. Simulations have demonstrated that the additive genetic model is robust even in cases of dominant or recessive inheritance [74, 199]. With this in mind, we assume that all loci fit the additive model. The expected value of S is calculated conditional on the parental genotypes under the assumption of Mendelian inheritance [74]. Under the null hypothesis, the FBAT statistic has an approximate standard normal distribution which can be used to calculate the significance of the observed test statistic.

A two-stage FBAT strategy was applied to correct for multiple testing in genome-wide and candidate gene association studies. Promising genetic loci are screened by ranking the power of the associated FBAT test statistics [198]. Significant associations are more likely to be identified at high powered loci in the downstream testing stage. To maintain independence between screening and testing, offspring genotypes in informative families (i.e. families in which at least one parent is heterozygous) are replaced by their expected value derived from parental genotypes. In this manner, parental genotypes and offspring phenotypes are used to estimate the genetic effect size, and subsequently the power, of the FBATs associated with each locus [198]. The second-stage uses the FBAT to evaluate observed offspring genotypes and identify loci with significant association with phenotype. Information obtained in the screening stage can be used to select significance thresholds and account for multiple testing. For example, testing the top *n* loci ranked in the screening stage greatly reduces the number of tests and increases the threshold at which genome-wide significance is implied (e.g. the top 10 loci can be tested with a significance threshold of 0.05/10) [174, 190]. A second strategy partitions the ranked loci into subsets of exponentially increasing size [200]. A significance threshold is defined for each subset by multiplying the genome-wide significance level (e.g. 0.05) with exponentially decreasing weights (Table 11). In this manner, all genetic loci can be tested for association without neglecting the information gained in the screening stage. In our analysis, we have adopted the latter method.

Table 11: Adjusted significance level determined by rank in screening step

Rank of Loci in Screening	Adjusted Significance Level
5	0.005
15	1.25E-03
35	3.12E-04
75	7.81E-05
155	1.95E-05
315	4.88E-06
635	1.22E-06
1,275	3.05E-07
2,555	7.63E-08
5,115	1.90E-08
10,235	4.77E-09
20,475	1.19E-09
40,955	2.98E-10

6.3.5 Assessment of linkage disequilibrium patterns

Linkage disequilibrium (LD) patterns in the multiplex cohort were analyzed to identify potential autism risk loci correlated with significant screened SNPs. Linkage disequilibrium is defined as the statistical association between two or more genetic loci and is calculated in a pairwise manner [201]. Consider two biallelic SNPs, in which two genetic variations are present at each locus, there are four potential haplotypes (i.e. snp1= A_1/B_1, snp2 = A_2/B_2, haplotypes = A_1A_2, A_1B_2, B_1A_2, B_1B_2). In the absence of LD, the expected frequency of the four haplotypes converges to the product of the constituent allele frequencies. Disequilibrium between the two loci can be calculated as:

$$D = P(A_1A_2) - P(A_1)P(A_2)$$

in which P(.) is the frequency of the corresponding haplotypes and alleles [202]. The *D* statistic is dependent on the allele frequencies in the population. A normalized measure of LD (*D'*) can be obtained by dividing *D* by its maximum value where

$$D_{\max} = \min(P(A_1)P(B_2), P(B_1)P(A_2))$$

The value of *D'* ranges from zero to one. Higher values correspond to higher disequilibrium. Pairs of genetic loci are said to be in strong LD if the one-sided upper 95% confidence bound of D' is greater than 0.98 and the lower bound is greater than 0.7 [201]. Conversely, loci with strong evidence of recombination can be identified by an upper confidence bound less than 0.7 [201]. A disequilibrium block is defined as a region over which less than 5% of pairwise comparisons show strong evidence of recombination [201]. The Haploview software package [203] was used to identify LD blocks in genotype data in the proximity of SNPs significantly associated with autism.

6.4 Results

6.4.1 Chromosomal location of candidate gene regions and SNPs tested for association

Candidate genes for autism-susceptibility loci were identified using gene expression analysis of an expression microarray dataset composed of 400 tissue samples and 19 different phenotypes. Neural-specific switch-like genes were identified with the following properties: the expression profile across all samples fits a bimodal distribution and expression is measured in the "high" mode in a majority of 89 samples of brain tissue. A set of 542 neural-specific switch-like genes were identified that meet these specifications. Figure 20 maps the chromosomal location of the coding and putative cis-regulatory regions of these genes as red bars to the left of the chromosome ideograms. Regulatory regions were conservatively identified using a 1 M base pair window centered at the midpoint of the gene coding region, as described

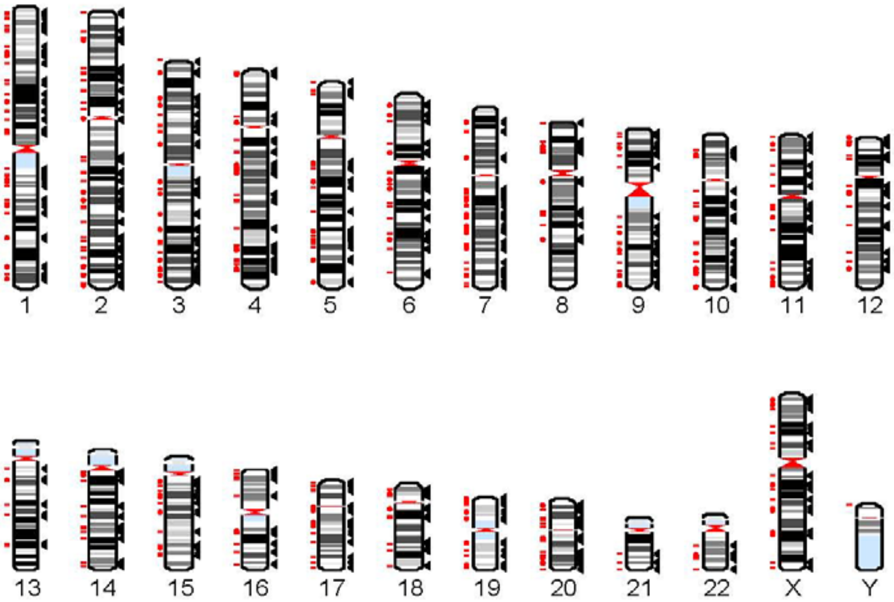


Figure 20: Karyogram depicting the chromosomal location of 542 neural-specific switch-like genes and 55,214 SNPs – Red bars to the left of the ideograms indicate the coding and putative regulatory regions of the identified neural-specific bimodal genes. Black arrows indicate the location of SNPs within these chromosomal regions.

[177]. Approximately 55,200 SNPs on the Affymetrix Genome-Wide Human SNP Array 5.0 platform are located in these chromosomal regions. Chromosomal locations of these SNPs are indicated in Figure 1 by black arrows. Candidate genes and associated SNPs are well-distributed throughout the autosomal and X chromosomes. Approximately 86% of the SNPs on the array were excluded through the use of the candidate gene approach.

6.4.2 SNP rs17101921 is an autism susceptibility locus

Association tests of SNPs in candidate gene regions identified an autism-susceptibility locus in the q26 region of chromosome 10. A two-stage FBAT strategy was used to test all of the loci in the candidate gene regions for association under an additive genetic model and correct for multiple testing. In the initial screening stage, loci were ranked according to estimates of statistical power of the corresponding association test. In the downstream testing stage, loci are tested for association with the FBAT statistic. The threshold for genome-wide significance for a genetic locus is determined by its rank in the screening stage (Table 11). Following this methodology, rs17101921 reached genome-wide significance (p-value = 0.0038; Table 12). Table 12 gives the rank in the screening stage, reference SNP identifier, minor allele frequency, FBAT p-value, genome-wide significance threshold and odds ratio for rs17101921 and adjacent SNPs on chromosome 10 represented on the Affymetrix chip. Individuals with the rs17101921 A allele are more likely to be autistic [odds ratio (OR) = 1.31, 95% confidence interval (CI) (0.81-2.11)]. In addition, none of the adjacent SNPs on the Affymetrix chip demonstrated significant association with autism susceptibility (Table 12). Association tests indicate that rs17101921 is a marker for autism risk.

6.4.3 Linkage disequilibrium analysis identifies a small haplotype block containing SNP rs17101921

We assessed linkage disequilibrium (LD) patterns around SNP rs17101921 to identify genetic variants that tend to be inherited with rs17101921 in haplotype blocks. Figure 21 shows a LD plot of SNPs on the Affymetrix chip in a ~125 kB region centered on the rs17101921 locus. The low LD between many pairs of loci indicates the rate of

Table 12: SNP rs17101921 associated with autism susceptibility- 47625 SNPs were tested for association with autism using a two-stage FBAT strategy. SNP rs17101921 in bold below passed the FBAT with genome-wide significance. Tested SNPs adjacent to rs17101921 were not significant. MAF = minor allele frequency, OR = odds ratio.

Rank in Screen	SNP	MAF	FBAT Pvalue	Significance Level	OR
6392	rs1896404	0.418	0.283	4.77E-09	0.9 (0.6-1.34)
470	rs2420929	0.131	0.2998	1.22E-06	2.51 (0.64-9.79)
5	**rs17101921**	**0.05**	**0.0038**	**0.005**	**1.31 (0.81-2.11)**
18895	rs9421422	0.485	0.4642	1.19E-09	0.84 (0.56-1.25)
9733	rs4457689	0.092	0.1514	4.77E-09	0.66 (0.21-2.02)

Table 13: Chromosomal location of rs17101921 and adjacent SNPs

SNP	Allele	Chromosome	Base Pair	Location
rs1896404	T/A	10	123131120	INTERGENIC
rs2420929	T/C	10	123143166	INTERGENIC
rs17101921	**G/A**	**10**	**123143285**	**INTERGENIC**
rs4457689	G/A	10	123166119	INTERGENIC
rs9421422	C/G	10	123144178	INTERGENIC

recombination in the region is high. SNP rs17101921 is located on a small LD block spanning ~13 kB consisting of four tested SNPs on the Affymetrix array, including rs9420328, rs1896404, rs2420929, and rs9421422. Although none of these SNPs demonstrated individual association with autism, testing the inheritance of the haplotype block as a whole may produce significant results. It should also be noted that rs17101921 may be in LD with SNPs that are not represented on the Affymetrix array. The dbSNP database identifies 97 SNPs in the same 13 kB region. Genotyping the chromosomal region at higher resolution will reveal a more detailed picture of linkage disequilibrium structure.

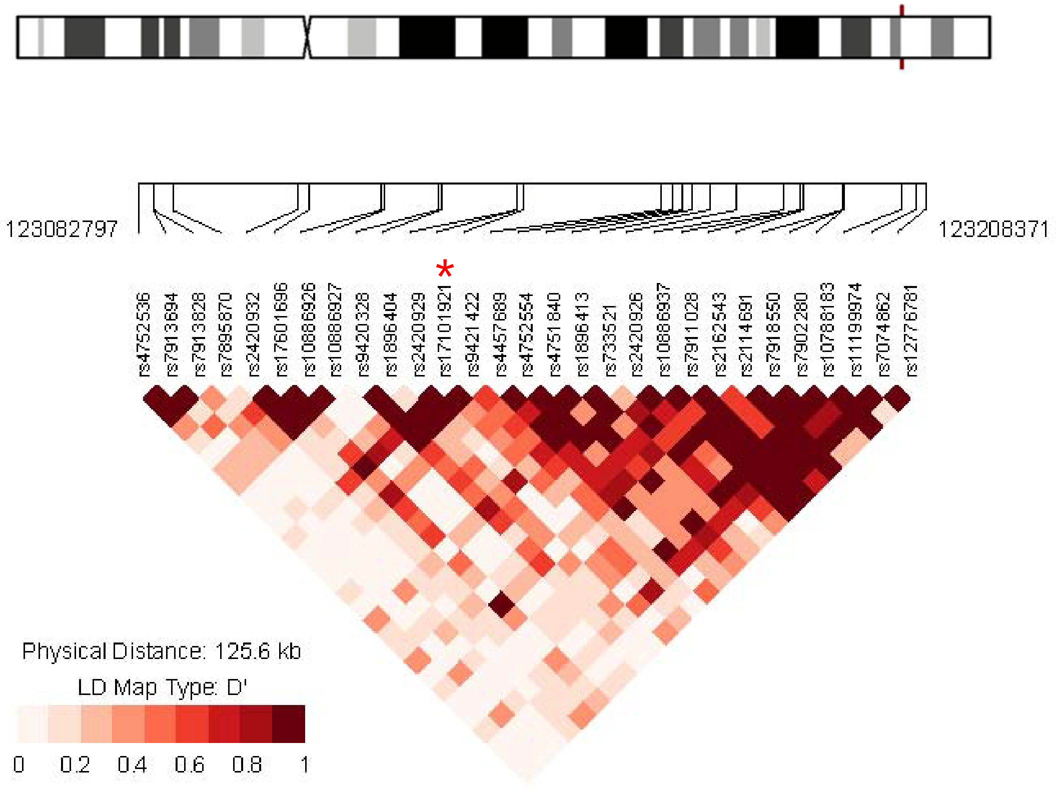


Figure 21: Linkage disequilibrium plot of the chromosomal region around rs17101921 – Pair-wise linkage disequilibrium between two SNPs is indicated by the color of the associated square along the diagonal. The red asterisk in the linkage disequilibrium plots identifies SNP rs17101921. Top figure: An ideogram of chromosome 10. The region mapped for linkage disequilibrium is indicated by the red bar. Bottom figure: A linkage disequilibrium plot of a ~125 kB region around SNP rs17101921. Linkage disequilibrium is quantified by D'.

6.4.4 Genes located within the q26 region of chromosome 10

To investigate potential functional implications of genetic variation at SNP rs17101921, we identified genes with coding regions located within the 1MB window around the genetic locus. The rs17101921 SNP is found in an intergenic region in the short arm of chromosome 10 (Table 13). Seven genes lie within this chromosomal region (Table 14). Table 14 lists the Entrez Gene identifiers, gene symbols, base pair locations of the start and end of the coding regions, and the identifier of the genes in the Online Mendelian Inheritance in Man (OMIM) database for these genes. Of the genes in the proximity of rs17101921, fibroblast growth factor receptor 2 (FGFR2) is the closest (~80 Kb downstream). The FGFR2 gene is also among the neural-specific switch-like genes used to identify candidate gene regions. These results suggest that autism susceptibility may be related to altered expression or function of FGFR2 as a result of genetic variation at rs17101921.

Table 14: Genes within 1Mb of rs17101921- Coding regions of seven genes are located within 1Mb of rs17101921. Fibroblast growth factor receptor 2 (FGFR2) in bold below is the closest to rs17101921 (~80Kb downstream) and is also a neural-specific bimodal gene.

Entrez Gene ID	Gene Symbol	Start	End	OMIM ID
196051	PPAPDC1A	122206456	122339357	
55717	BRWD2	122600860	122659025	606417
2263	**FGFR2**	**123223889**	**123347962**	**176943**
11101	ATE1	123492616	123677936	607103
54780	NSMCE4A	123706601	123724722	
10579	TACC2	123738699	124004049	605302
118663	BTBD16	124020811	124087666	

6.5 Discussion

In this study, we have found a significant association between autism and genetic variation at a single genetic locus in an intergenic region of chromosome 10 in a cohort of 189 multiplex autism-affected families. Several measures were adopted to reduce the severity of multiple testing corrections applied to family-based association tests. In previous work, we identified a set of neural-specific switch-like genes with bimodal expression patterns and known involvement in nervous system development and function. Using the coding and cis-regulatory regions of these genes as candidate association loci resulted in the reduction of the SNP feature space by 86%. A two-stage family-based association test was used for association testing [200]. In this procedure, SNPs are first ranked according to the probability that a significant association will be found and then tested for association with the significance threshold dependent on the rank in the screening stage. With this method, SNP rs17101921 was identified as an autism-susceptibility locus with genome-wide significance (p-value = 0.0038; threshold = 0.005). Individuals with the A allele at this locus were found to be at greater risk for developing autism [OR = 1.31, CI (0.81-2.11)]. Linkage disequilibrium analysis localizes rs17101921 to a small haplotype block consisting of four other SNPs on the Affymetrix array (rs9420328, rs1896404, rs2420929, and rs9421422). In addition, rs17101921 is located approximately 80 kB upstream from FGFR2, a growth factor receptor involved in neurodevelopment and neural function.

Linkage between genomic regions on chromosome 10 and neurological disorders has been observed in a number of previously published studies. A significant (p-value < 0.01) quantitative trait locus linked to social responsiveness scores was identified in a genomic screen of 62 families with male autistic children [204]. Similarly, a quantitative trait locus linked to a measure of language development, age at first phrase, was detected (p-value = 0.018) in a study of 152 multiplex families [205]. A third study identified autism-linked genetic loci at two locations along chromosome 10 (10p14; 10q23.31; p-values not given) in an analysis of affected families showing elevated obsessive-compulsive traits [206]. A meta-analysis of data from five independent genome scans identified several loci (10p12-q11.1, p-value = 0.0022; 10q11.2-q23, p-value = 0.0299; 10q22-q23, p-value = 0.0432) linked with nominal significance to risk for autism-spectrum disorders [184]. In addition, linkage has been reported between the 10q26 region and a number of other neurological disorders, including schizophrenia [207, 208], bipolar disorder [208] and Alzheimer's disease [209]. These studies, along with our findings, provide evidence of the involvement of genetic loci on chromosome 10 in conferring autism susceptibility and suggest the possibility of common molecular mechanisms in autism, schizophrenia, bipolar disorder and Alzheimer's disease.

Genetic variation at rs17101921 may influence the expression of fibroblast growth factor receptor 2 (FGFR2) located approximately 80 kB downstream. Genomic screens for expression quantitative trait loci (eQTL), have established that genetic variants up to 100 kB from the coding region can influence gene expression [176, 177]. Fibroblast growth factor receptors are transmembrane proteins with an intracellular tyrosine kinase domain. Ligand binding induces dimerization and receptor activation by phosphorylation of the intracellular domain [192]. Signaling through fibroblast growth factor receptors activates a number of downstream processes including proliferation, cell-cycle progression and cytoskeletal remodeling[192]. Fibroblast growth factor receptor 2 (FGFR2) is highly expressed in glial cells in the brain [210]. In both the developing and adult brain, expression of fibroblast growth factor-2 (FGF2) increases the proliferation of neurons and neural stem cells, stimulates axon branching, and has a neuroprotective function in brain injury and ischemia [192]. Larger brain size at age 2-4, increased glial cell activation and excessive neural degeneration later in life are all characteristics of autism

[211, 212]. Altered regulation of FGFR2 expression could be a contributing factor in all of these conditions. In addition, mutation of the FGFR2 gene has been associated with a number of developmental disorders including craniosyntosis and Crouzon syndrome which presents with mental disabilities [213]. These findings suggest that increased autism risk transmitted by variation at the rs17101921 locus is potentially related to expression of the FGFR2 gene.

6.6 Conclusion

We detected an autism susceptibility locus, SNP rs17101921, through family-based association testing of genomic data obtained from a cohort of 189 mulitplex families in the AGRE database. Candidate gene regions were identified using neural-specific bimodal expression patterns identified from gene expression analysis of a large, phenotypically diverse compilation of microarray data. The positive results of this approach validate future study of genetic variation in candidate gene regions identified from microarray analysis. Functional characterization of the genetic variant is frustrated by localization to an intergenic chromosomal region; however, a potential link is established through the FGFR2 gene. Analysis of gene expression data linked to genotype information could be used to verify this connection. Results justify further analysis of the 10q26 chromosomal region with genotyping at higher resolution. In addition, a screen for quantitative trait loci associated with traits such as social responsiveness, language development or repetitive behaviors may result in a genomic screen with greater statistical power.

Chapter 7: Conclusion

Technological advances in the post-genomic era of biological study have provided a higher resolution picture of gene regulation and the perturbations that characterize complex phenomena such as development, differentiation and chronic disease. High-throughput sequencing platforms and gene expression microarrays provide the means to profile biological systems from the genomic and transcriptomic perspectives. In addition, shared public databases catalog the sequence, structure and function of genes and proteins and organize them into coherent ontologies and interaction models. The abundance of data available allows for an unbiased, holistic approach to investigation. Significant results can be used to gain new insight and generate original testable hypotheses. Conversely, the high dimensionality of genome-scale datasets makes it difficult to identify relevant information from noise. In this work, we integrated prior knowledge, gene functional information, and genome-scale microarray data across two modalities with classification methodologies to extract meaningful biological information from high dimensional microarray datasets.

In the study of cancer, gene expression microarray analysis has been used to identify expression biomarkers that either classify samples into clinically homogenous subtypes or predict the course of disease. With extensive use of this methodology, it was observed that multiple biomarker sets generated from independent studies of the same disease state share few common genes. We developed an iterative supervised classification approach to generate populations of biomarker sets that could be evaluated for predictive potential. Results indicate that many biomarker sets accurately classify cancer samples. A possible biological explanation for the lack of agreement between biomarker panels is the redundancy observed in cellular signaling pathways. This explanation is supported by analyses that demonstrate more consistent results when comparing independent datasets by the expression of functional sets of genes rather than individual genes [92]. Technical variability between microarray platforms may also play a role. To investigate this possibility, we assessed the classification accuracy of expression biomarker panels on independent datasets both within and across microarray platforms. Predictive accuracy decreased and was more variable when panels were tested across platforms, indicating that technical variability is a significant issue. Studies analyzing the reproducibility of microarray expression values in multi-center trials have indicated that technical variability can be adequately controlled by proper experimental design [214].

Large-scale compilation of gene expression datasets in public repositories provides the opportunity to investigate patterns of gene expression across diverse biological phenotypes. A number of studies have used gene expression analysis to identify house-keeping genes ubiquitously expressed across different tissue types and presumably required for normal cellular function [132, 133]. Similarly, switch-like genes with bimodal expression profiles have been identified [134, 144]. We used a number of classification methods to investigate the expression of switch-like genes in datasets of diverse phenotypes in health and disease. Use of a model-based classification method resulted in accurate classification of tissues into groups corresponding to 19 different tissue types. Similar accuracy was obtained with a multi-class supervised classification method. These results suggest that identification of switch-like genes may be an effective method to reduce the size of the feature space in gene expression analysis. Model-based and supervised methods also produced accurate classification of samples in a dataset profiling blood cells from subjects with one of four infectious diseases. A number of studies have observed unique expression changes in peripheral blood cells of the immune system in response to different pathogens [215, 216]. The biological relevance of bimodal expression patterns is implied by functional enrichment analysis of the activated switch-like genes in different phenotypes. In tissues including the brain, and skeletal and cardiac muscle, activated switch-like genes are enriched for tissue-specific functions. Similarly, in infectious disease activated genes are enriched with functions related to the immune response. In light of these results, switch-like genes appear to be involved in specialized or temporally active biological processes. Identification and characterization of switch-like genes as well as the phenotypes in which they are activated will have important implications in fields such as stem cell research, tissue engineering and gene therapy.

Genetic variation at the gene sequence level can result in differences in the expression and function of gene products that contribute to increased risk of disease. Single nucleotide polymorphisms have been associated with a number of common pathological conditions including obesity [174], diabetes [34, 175], inflammatory bowel disease [34, 173], and cardiovascular disease [34, 172]. Using information gained from analysis of switch-like expression patterns, we detected a single nucleotide polymorphism in an intergenic region of chromosome 10 associated with increased susceptibility to autism. Notably, the same chromosomal region has been linked to other neurological diseases including bipolar disorder [208], schizophrenia [207, 208] and Alzheimier's disease [209]. The functional implications of genetic variation in gene coding regions are assessed by examining sequence, structure and evolutionary conservation of homologous proteins [75, 76]. Less is known about the effects of variation in non-coding regions; however the increased autism risk transmitted by the genetic variant identified in our analysis may be related to altered expression of the fibroblast growth factor receptor 2 gene. Analysis of genomic and transcriptomic data from a set of autistic individuals to validate this hypothesis is warranted. In addition, identification of the autism susceptibility locus in this analysis motivates the development of more direct methods for the integration of gene expression and gene mapping microarray datasets.

List of References

1. Loden M, van Steensel B: **Whole-genome views of chromatin structure.** *Chromosome Res* 2005, **13:**289-298.

2. Rothkamm K, Gunasekara K, Warda SA, Krempler A, Lobrich M: **Radiation-induced HPRT mutations resulting from misrejoined DNA double-strand breaks.** *Radiat Res* 2008, **169:**639-648.

3. Thomas HC, Foster GR, Sumiya M, McIntosh D, Jack DL, Turner MW, Summerfield JA: **Mutation of gene of mannose-binding protein associated with chronic hepatitis B viral infection.** *Lancet* 1996, **348:**1417-1419.

4. Pollard TD, Earnshaw WC: *Cell Biology.* Philadelphia: Elsevier 2004.

5. Wu X, Tu X, Joeng KS, Hilton MJ, Williams DA, Long F: **Rac1 activation controls nuclear localization of beta-catenin during canonical Wnt signaling.** *Cell* 2008, **133:**340-353.

6. Veselik DJ, Divekar S, Dakshanamurthy S, Storchan GB, Turner JM, Graham KL, Huang L, Stoica A, Martin MB: **Activation of estrogen receptor-alpha by the anion nitrite.** *Cancer Res* 2008, **68:**3950-3958.

7. Wenta N, Strauss H, Meyer S, Vinkemeier U: **Tyrosine phosphorylation regulates the partitioning of STAT1 between different dimer conformations.** *Proc Natl Acad Sci U S A* 2008, **105:**9238-9243.

8. Holmberg CI, Tran SE, Eriksson JE, Sistonen L: **Multisite phosphorylation provides sophisticated regulation of transcription factors.** *Trends Biochem Sci* 2002, **27:**619-627.

9. Modrek B, Lee C: **A genomic view of alternative splicing.** *Nat Genet* 2002, **30:**13-19.

10. **Lawrence Berkeley National Laboratory Image Library** [http://acs.lbl.gov/ImgLib/COLLECTIONS/BERKELEY-LAB/RESEARCH-1991-PRESENT/LIFE-SCIENCES/index/96703355.html]

11. Walsh CT, Garneau-Tsodikova S, Gatto GJ, Jr.: **Protein posttranslational modifications: the chemistry of proteome diversifications.** *Angew Chem Int Ed Engl* 2005, **44:**7342-7372.

12. Vadaie N, Dionne H, Akajagbor DS, Nickerson SR, Krysan DJ, Cullen PJ: **Cleavage of the signaling mucin Msb2 by the aspartyl protease Yps1 is required for MAPK activation in yeast.** *J Cell Biol* 2008, **181:**1073-1081.

13. Bird A: **Perceptions of epigenetics.** *Nature* 2007, **447:**396-398.

14. Jaenisch R, Bird A: **Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.** *Nat Genet* 2003, **33 Suppl:**245-254.

15. Saxonov S, Berg P, Brutlag DL: **A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.** *Proc Natl Acad Sci U S A* 2006, **103:**1412-1417.

16. Geiman TM, Robertson KD: **Chromatin remodeling, histone modifications, and DNA methylation-how does it all fit together?** *J Cell Biochem* 2002, **87:**117-125.

17. Vignali M, Hassan AH, Neely KE, Workman JL: **ATP-dependent chromatin-remodeling complexes.** *Mol Cell Biol* 2000, **20:**1899-1910.

18. Shyu AB, Wilkinson MF, van Hoof A: **Messenger RNA regulation: to translate or to degrade.** *Embo J* 2008, **27:**471-481.

19. Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, Zamore PD: **A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA.** *Science* 2001, **293:**834-838.

20. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN: **The nuclear RNase III Drosha initiates microRNA processing.** *Nature* 2003, **425:**415-419.

21. Peters L, Meister G: **Argonaute proteins: mediators of RNA silencing.** *Mol Cell* 2007, **26:**611-623.

22. Vasudevan S, Tong Y, Steitz JA: **Switching from repression to activation: microRNAs can up-regulate translation.** *Science* 2007, **318:**1931-1934.

23. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449:**851-861.

24. **A haplotype map of the human genome.** *Nature* 2005, **437:**1299-1320.

25. Garrington TP, Johnson GL: **Organization and regulation of mitogen-activated protein kinase signaling pathways.** *Curr Opin Cell Biol* 1999, **11:**211-218.

26. Johnson GL, Lapadat R: **Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases.** *Science* 2002, **298:**1911-1912.

27. Pimienta G, Pascual J: **Canonical and alternative MAPK signaling.** *Cell Cycle* 2007, **6:**2628-2632.

28. Junttila MR, Li SP, Westermarck J: **Phosphatase-mediated crosstalk between MAPK signaling pathways in the regulation of cell survival.** *Faseb J* 2008, **22:**954-965.

29. Shaul YD, Seger R: **The MEK/ERK cascade: from signaling specificity to diverse functions.** *Biochim Biophys Acta* 2007, **1773:**1213-1226.

30. Lawrence MC, Jivan A, Shao C, Duan L, Goad D, Zaganjor E, Osborne J, McGlynn K, Stippec S, Earnest S, et al: **The roles of MAPKs in disease.** *Cell Res* 2008, **18:**436-442.

31. Esteller M: **Epigenetic gene silencing in cancer: the DNA hypermethylome.** *Hum Mol Genet* 2007, **16 Spec No 1:**R50-59.

32. Badano JL, Katsanis N: **Beyond Mendel: an evolving view of human genetic disease transmission.** *Nat Rev Genet* 2002, **3:**779-789.

33. Botstein D, Risch N: **Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.** *Nat Genet* 2003, **33 Suppl:**228-237.

34. **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661-678.

35. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: **Expression profiling using cDNA microarrays.** *Nat Genet* 1999, **21:**10-14.

36. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays.** *Nat Genet* 1999, **21:**20-24.

37. Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, et al: **Large-scale genotyping of complex DNA.** *Nat Biotechnol* 2003, **21:**1233-1237.

38. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet* 2001, **29:**365-371.

39. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, et al: **An overview of Ensembl.** *Genome Res* 2004, **14:**925-928.

40. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2005, **33:**D54-58.

41. **The universal protein resource (UniProt).** *Nucleic Acids Res* 2008, **36:**D190-195.

42. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30:**207-210.

43. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, et al: **ArrayExpress--a public database of microarray experiments and gene expression profiles.** *Nucleic Acids Res* 2007, **35:**D747-750.

44. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.

45. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.

46. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4:**249-264.

47. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, **31:**e15.

48. **Statistical algorithms description document** [http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf]

49. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19:**185-193.

50. Katz S, Irizarry RA, Lin X, Tripputi M, Porter MW: **A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database.** *BMC Bioinformatics* 2006, **7:**464.

51. **BRLMM: an Improved Genotype Calling Method for the GeneChip Human Mapping 500K Array Set**

52. Mariani TJ, Budhraja V, Mecham BH, Gu CC, Watson MA, Sadovsky Y: **A variable fold change threshold determines significance for expression microarrays.** *Faseb J* 2003, **17:**321-323.

53. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci U S A* 2001, **98:**5116-5121.

54. Kerr MK, Martin M, Churchill GA: **Analysis of variance for gene expression microarray data.** *J Comput Biol* 2000, **7:**819-837.

55. Schultz IJ, Wester K, Straatman H, Kiemeney LA, Babjuk M, Mares J, Willems JL, Swinkels DW, Witjes JA, Malmstrom PU, de Kok JB: **Gene expression analysis for the prediction of recurrence in patients with primary Ta urothelial cell carcinoma.** *Eur Urol* 2007, **51:**416-422; discussion 422-413.

56. Bland JM, Altman DG: **Multiple significance tests: the Bonferroni method.** *Bmj* 1995, **310:**170.

57. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society Series B-Methodological* 1995, **57:**289-300.

58. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.

59. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.

60. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, et al: **Repeated observation of breast tumor subtypes in independent gene expression data sets.** *Proc Natl Acad Sci U S A* 2003, **100:**8418-8423.

61. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95:**14863-14868.

62. Medvedovic M, Sivaganesan S: **Bayesian infinite mixture model based clustering of gene expression profiles.** *Bioinformatics* 2002, **18:**1194-1206.

63. Ghosh D, Chinnaiyan AM: **Mixture modelling of gene expression data from microarray experiments.** *Bioinformatics* 2002, **18:**275-286.

64. McLachlan GJ, Bean RW, Peel D: **A mixture model-based approach to the clustering of microarray expression data.** *Bioinformatics* 2002, **18:**413-422.

65. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: **Model-based clustering and data transformations for gene expression data.** *Bioinformatics* 2001, **17:**977-987.

66. Joshi A, Van de Peer Y, Michoel T: **Analysis of a Gibbs sampler method for model-based clustering of gene expression data.** *Bioinformatics* 2008, **24:**176-183.

67. Qin ZS: **Clustering microarray gene expression data using weighted Chinese restaurant process.** *Bioinformatics* 2006, **22:**1988-1997.

68. Simon R: **Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data.** *Br J Cancer* 2003, **89:**1599-1604.

69. Zhang B, Kirov S, Snoddy J: **WebGestalt: an integrated system for exploring gene sets in various biological contexts.** *Nucleic Acids Res* 2005, **33:**W741-748.

70. Dawn Teare M, Barrett JH: **Genetic linkage studies.** *Lancet* 2005, **366:**1036-1044.

71. Pritchard JK, Donnelly P: **Case-control studies of association in structured or admixed populations.** *Theor Popul Biol* 2001, **60:**227-237.

72. Devlin B, Roeder K, Wasserman L: **Genomic control, a new approach to genetic-based association studies.** *Theor Popul Biol* 2001, **60:**155-166.

73. Satten GA, Flanders WD, Yang Q: **Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model.** *Am J Hum Genet* 2001, **68:**466-477.

74. Horvath S, Xu X, Laird NM: **The family based association test method: strategies for studying general genotype--phenotype associations.** *Eur J Hum Genet* 2001, **9:**301-306.

75. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31:**3812-3814.

76. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey.** *Nucleic Acids Res* 2002, **30:**3894-3900.

77. Chen YW, Zhao P, Borup R, Hoffman EP: **Expression profiling in the muscular dystrophies: identification of novel aspects of molecular pathophysiology.** *J Cell Biol* 2000, **151:**1321-1336.

78. Puricelli L, Iori E, Millioni R, Arrigoni G, James P, Vedovato M, Tessari P: **Proteome analysis of cultured fibroblasts from type 1 diabetic patients and normal subjects.** *J Clin Endocrinol Metab* 2006, **91:**3507-3514.

79. Barnes MG, Aronow BJ, Luyrink LK, Moroldo MB, Pavlidis P, Passo MH, Grom AA, Hirsch R, Giannini EH, Colbert RA, et al: **Gene expression in juvenile arthritis and spondyloarthropathy: pro-angiogenic ELR+ chemokine genes relate to course of arthritis.** *Rheumatology (Oxford)* 2004, **43:**973-979.

80. Ma J, Liew CC: **Gene profiling identifies secreted protein transcripts from peripheral blood cells in coronary artery disease.** *J Mol Cell Cardiol* 2003, **35:**993-998.

81. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, et al: **Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nat Med* 2002, **8:**68-74.

82. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, et al: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347:**1999-2009.

83. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, et al: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415:**530-536.

84. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA, Jr., Marks JR, Nevins JR: **Predicting the clinical status of human breast cancer by using gene expression profiles.** *Proc Natl Acad Sci U S A* 2001, **98:**11462-11467.

85. Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data.** *Journal of the American Statistical Association* 2002, **97:**77-87.

86. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z: **Tissue classification with gene expression profiles.** *Journal of Computational Biology* 2000, **7:**559-583.

87. Guyon I, Weston J, Barnhill S: **Gene selection for cancer classification using support vector machines.** *Machine Learning* 2002, **46:**389-422.

88. Li L, Darden TA, Weinberg CR, Levine AJ, Pedersen LG: **Gene assessment and sample classification for gene expression data using agenetic algorithm/k-nearest neighbor method.** *Combinatorial Chemistry and High Throughput Screening* 2001, **4:**727-739.

89. Liu JJ, Cutler G, Li W, Pan Z, Peng S, Hoey T, Chen L, Ling X: **Multiclass cancer classification and biomarker discovery using GA-based algorithms** *Bioinformatics* 2005, **21:**2691-2697.

90. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer: is there a unique set?** *Bioinformatics* 2005, **21:**171-178.

91. Fortunel NO, Otu HH, Ng HH, Chen J, Mu X, Chevassut T, Li X, Joseph M, Bailey C, Hatzfeld JA, et al: **Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature".** *Science* 2003, **302:**393; author reply 393.

92. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.

93. Wang H, He X, Band M, Wilson C, Liu L: **A study of inter-lab and inter-platform agreement of DNA microarray data.** *BMC Genomics* 2005, **6:**71.

94. Zakharkin SO, Kim K, Mehta T, Chen L, Barnes S, Scheirer KE, Parrish RS, Allison DB, Page GP: **Sources of variation in Affymetrix microarray experiments.** *BMC Bioinformatics* 2005, **6:**214.

95. Baker SG, Kramer BS: **Identifying genes that contribute most to good classification in microarrays.** *BMC Bioinformatics* 2006, **7:**407.

96. Grate LR: **Many accurate small-discriminatory feature subsets exist in microarray transcript data: biomarker discovery.** *BMC Bioinformatics* 2005, **6:**97.

97. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy.** *Lancet* 2005, **365:**488-492.

98. Hummel M, Bentink S, Berger H, Klapper W, Wessendorf S, Barth TF, Bernd HW, Cogliatti SB, Dierlamm J, Feller AC, et al: **A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling.** *N Engl J Med* 2006, **354:**2419-2430.

99. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J: **An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival (vol 102, pg 13550, 2005).** *Proc Natl Acad Sci U S A* 2005, **102:**17882-17882.

100. Monti S, Savage KJ, Kutok JL, Feuerhake F, Kurtin P, Mihm M, Wu B, Pasqualucci L, Neuberg D, Aguiar RC, et al: **Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response.** *Blood* 2005, **105:**1851-1861.

101. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, et al: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *Lancet* 2005, **365:**671-679.

102. Zhao H, Ljungberg B, Grankvist K, Rasmuson T, Tibshirani R, Brooks JD: **Gene expression profiling predicts survival in conventional renal cell carcinoma.** *PLoS Med* 2006, **3:**e13.

103. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, et al: **The Stanford Microarray Database.** *Nucleic Acids Res* 2001, **29:**152-155.

104. Ihaka R, Gentleman R: **R: a language for data analysis and graphics.** *J Comput Graph Stat* 1996, **3:**299-314.

105. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5:**R80.

106. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17:**520-525.

107. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA: **SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.** *Nucleic Acids Res* 2003, **31:**219-223.

108. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, Rapp BA: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2000, **28:**10-14.

109. **GeneChip expression data analysis fundamentals.**

110. Molinaro AM, Simon R, Pfeiffer RM: **Prediction error estimation: a comparison of resampling methods.** *Bioinformatics* 2005, **21:**3301-3307.

111. **Ingenuity Pathways Analysis** [http://www.ingenuity.com/]

112. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L: **The use of receiver operating characteristic curves in biomedical informatics.** *J Biomed Inform* 2005, **38:**404-415.

113. Macskassy S, Provost R, Rosset S: **Confidence bands for ROC curves: methods and an empirical study** In *Proceedings of the 22nd Internationl Conference on Machine Learning; Bonn, Germany* 2005

114. Bura E, Pfeiffer RM: **Graphical methods for class prediction using dimension reduction techniques on DNA microarray data.** *Bioinformatics* 2003, **19:**1252-1258.

115. Reiner A, Yekutieli D, Benjamini Y: **Identifying differentially expressed genes using false discovery rate controlling procedures.** *Bioinformatics* 2003, **19:**368-375.

116. Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C, Meltzer PS: **Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns.** *Cancer Res* 2001, **61:**5979-5984.

117. Bjornstrom L, Sjoberg M: **Mechanisms of estrogen receptor signaling: convergence of genomic and nongenomic actions on target genes.** *Mol Endocrinol* 2005, **19:**833-842.

118. Chang CC, Ye BH, Chaganti RS, Dalla-Favera R: **BCL-6, a POZ/zinc-finger protein, is a sequence-specific transcriptional repressor.** *Proc Natl Acad Sci U S A* 1996, **93:**6947-6952.

119. Cattoretti G, Chang CC, Cechova K, Zhang J, Ye BH, Falini B, Louie DC, Offit K, Chaganti RS, Dalla-Favera R: **BCL-6 protein is expressed in germinal-center B cells.** *Blood* 1995, **86:**45-53.

120. Hans CP, Weisenburger DD, Greiner TC, Gascoyne RD, Delabie J, Ott G, Muller-Hermelink HK, Campo E, Braziel RM, Jaffe ES, et al: **Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray.** *Blood* 2004, **103:**275-282.

121. Lossos IS, Jones CD, Warnke R, Natkunam Y, Kaizer H, Zehnder JL, Tibshirani R, Levy R: **Expression of a single gene, BCL-6, strongly predicts survival in patients with diffuse large B-cell lymphoma.** *Blood* 2001, **98:**945-951.

122. Bland JM, Altman DG: **Survival probabilities (the Kaplan-Meier method).** *Bmj* 1998, **317:**1572.

123. Mecham BH, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, Wetmore DZ, Mariani TJ, Kohane IS, Szallasi Z: **Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements.** *Nucleic Acids Res* 2004, **32:**e74.

124. Choi JK, Yu U, Kim S, Yoo OJ: **Combining multiple microarray studies and modeling interstudy variation.** *Bioinformatics* 2003, **19 Suppl 1:**i84-90.

125. Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E: **A cross-study comparison of gene expression studies for the molecular classification of lung cancer.** *Clin Cancer Res* 2004, **10:**2922-2927.

126. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: **Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer.** *Cancer Res* 2002, **62:**4427-4433.

127. Li L, Chen L, Goldgof D, George F, Chen Z, Rao A, Cragun J, Sutphen R, Lancaster J: **Integration of clinical information and gene expression profiles for prediction of chemo-response for ovarian cancer.** *Conf Proc IEEE Eng Med Biol Soc* 2005, **5:**4818-4821.

128. Pittman J, Huang E, Dressman H, Horng CF, Cheng SH, Tsou MH, Chen CM, Bild A, Iversen ES, Huang AT, et al: **Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes.** *Proc Natl Acad Sci U S A* 2004, **101:**8431-8436.

129. Sun Y, Goodison S, Li J, Liu L, Farmerie W: **Improved breast cancer prognosis through the combination of clinical and genetic markers.** *Bioinformatics* 2007, **23:**30-37.

130. Arora A, Simpson DA: **Individual mRNA expression profiles reveal the effects of specific microRNAs.** *Genome Biol* 2008, **9:**R82.

131. Hobert O: **Gene regulation by transcription factors and microRNAs.** *Science* 2008, **319:**1785-1786.

132. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, et al: **A compendium of gene expression in normal human tissues.** *Physiol Genomics* 2001, **7:**97-104.

133. Warrington JA, Nair A, Mahadevappa M, Tsyganskaya M: **Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes.** *Physiol Genomics* 2000, **2:**143-147.

134. Ertel A, Tozeren A: **Human switch-like genes and their regulation via transcription initiation and histone methylation.** *BMC Genomics* 2008.

135. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, et al: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403:**503-511.

136. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proc Natl Acad Sci U S A* 1999, **96:**6745-6750.

137. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285.

138. Toronen P, Kolehmainen M, Wong G, Castren E: **Analysis of gene expression data using self-organizing maps.** *FEBS Lett* 1999, **451:**142-146.

139. Medvedovic M, Yeung KY, Bumgarner RE: **Bayesian mixture model based clustering of replicated microarray data.** *Bioinformatics* 2004, **20:**1222-1232.

140. Fan J, May SJ, Zhou Y, Barrett-Connor E: **Bimodality of 2-h plasma glucose distributions in whites: the Rancho Bernardo study.** *Diabetes Care* 2005, **28:**1451-1456.

141. Lim TO, Bakri R, Morad Z, Hamid MA: **Bimodality in blood glucose distribution: is it universal?** *Diabetes Care* 2002, **25:**2212-2217.

142. Maclean CJ, Morton NE, Elston RC, Yee S: **Skewness in commingled distributions.** *Biometrics* 1976, **32:**695-699.

143. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood from incomplete data via the EM alogrithm** *Journal of the Royal Statistical Society* 1977, **39:**1-38.

144. Ertel A, Tozeren A: **Switch-like genes populate cell communication pathways and are enriched for extracellular proteins.** *BMC Genomics* 2008, **9:**3.

145. Hartigan JA, Wong MA: **A K-means clustering algorithm** *Applied Statistics* 1979, **28:**100-108.

146. Tibshirani R, Walther G, Hastie T: **Estimating the number of clusters in a data set via the gap statistic.** *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2001, **63:**411-423.

147. Hoff PD: **Model-based subspace clustering.** *Bayesian Analysis* 2006, **1:**321-344.

148. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, et al: **Multiclass cancer diagnosis using tumor gene expression signatures.** *Proc Natl Acad Sci U S A* 2001, **98:**15149-15154.

149. Hynes RO: **Integrins: bidirectional, allosteric signaling machines.** *Cell* 2002, **110:**673-687.

150. Wheelock MJ, Johnson KR: **Cadherin-mediated cellular signaling.** *Curr Opin Cell Biol* 2003, **15:**509-514.

151. De Arcangelis A, Georges-Labouesse E: **Integrin and ECM functions: roles in vertebrate development.** *Trends Genet* 2000, **16:**389-395.

152. Nelson CM, Bissell MJ: **Of extracellular matrix, scaffolds, and signaling: tissue architecture regulates development, homeostasis, and cancer.** *Annu Rev Cell Dev Biol* 2006, **22:**287-309.

153. Bon G, Folgiero V, Di Carlo S, Sacchi A, Falcioni R: **Involvement of alpha6beta4 integrin in the mechanisms that regulate breast cancer progression.** *Breast Cancer Res* 2007, **9:**203.

154. Buttery RC, Rintoul RC, Sethi T: **Small cell lung cancer: the importance of the extracellular matrix.** *Int J Biochem Cell Biol* 2004, **36:**1154-1160.

155. van Horssen J, Dijkstra CD, de Vries HE: **The extracellular matrix in multiple sclerosis pathology.** *J Neurochem* 2007, **103:**1293-1301.

156. Gormley M, Dampier W, Ertel A, Karacali B, Tozeren A: **Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets.** *BMC Bioinformatics* 2007, **8:**415.

157. Yamada S, Nelson WJ: **Synapses: sites of cell recognition, adhesion, and functional specification.** *Annu Rev Biochem* 2007, **76:**267-294.

158. Scherer SS, Arroyo EJ: **Recent progress on the molecular organization of myelinated axons.** *J Peripher Nerv Syst* 2002, **7:**1-12.

159. Janeway CA, Jr., Medzhitov R: **Innate immune recognition.** *Annu Rev Immunol* 2002, **20:**197-216.

160. Pasare C, Medzhitov R: **Toll-like receptors: linking innate and adaptive immunity.** *Microbes Infect* 2004, **6:**1382-1387.

161. Guida M, D'Elia G, Benvestito S, Casamassima A, Micelli G, Quaranta M, Moschetta R, De Lena M, Lorusso V: **Hepatitis C virus infection in patients with B-cell lymphoproliferative disorders.** *Leukemia* 2002, **16:**2162-2163.

162. Landau DA, Saadoun D, Calabrese LH, Cacoub P: **The pathophysiology of HCV induced B-cell clonal disorders.** *Autoimmun Rev* 2007, **6:**581-587.

163. Lindenschmidt EG, Granato CH, Katzner K, Laufs R: **Evidence for limited humoral immunoglobulin M antibody response to hepatitis B core antigen during acute and chronic hepatitis B virus infections.** *J Clin Microbiol* 1985, **21:**1000-1003.

164. Bureau C, Bernad J, Chaouche N, Orfila C, Beraud M, Gonindard C, Alric L, Vinel JP, Pipy B: **Nonstructural 3 protein of hepatitis C virus triggers an oxidative burst in human monocytes via activation of NADPH oxidase.** *J Biol Chem* 2001, **276:**23077-23083.

165. Sarantis H, Gray-Owen SD: **The specific innate immune receptor CEACAM3 triggers neutrophil bactericidal activities via a Syk kinase-dependent pathway.** *Cell Microbiol* 2007, **9:**2167-2180.

166. Anand AR, Ganju RK: **HIV-1 gp120-mediated apoptosis of T cells is regulated by the membrane tyrosine phosphatase CD45.** *J Biol Chem* 2006, **281:**12289-12299.

167. Barat C, Tremblay MJ: **Engagement of CD43 enhances human immunodeficiency virus type 1 transcriptional activity and virus production that is induced upon TCR/CD3 stimulation.** *J Biol Chem* 2002, **277:**28714-28724.

168. Perfettini JL, Roumier T, Castedo M, Larochette N, Boya P, Raynal B, Lazar V, Ciccosanti F, Nardacci R, Penninger J, et al: **NF-kappaB and p53 are the dominant apoptosis-inducing transcription factors elicited by the HIV-1 envelope.** *J Exp Med* 2004, **199:**629-640.

169. Shyamsundar R, Kim YH, Higgins JP, Montgomery K, Jorden M, Sethuraman A, van de Rijn M, Botstein D, Brown PO, Pollack JR: **A DNA microarray survey of gene expression in normal human tissues.** *Genome Biol* 2005, **6:**R22.

170. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO: **Individuality and variation in gene expression patterns in human blood.** *Proc Natl Acad Sci U S A* 2003, **100:**1896-1901.

171. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al: **Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.** *Bioinformatics* 2005, **21:**650-659.

172. Arking DE, Pfeufer A, Post W, Kao WH, Newton-Cheh C, Ikeda M, West K, Kashuk C, Akyol M, Perz S, et al: **A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization.** *Nat Genet* 2006, **38:**644-651.

173. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, et al: **A genome-wide association study identifies IL23R as an inflammatory bowel disease gene.** *Science* 2006, **314:**1461-1463.

174. Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, et al: **A common genetic variant is associated with adult and childhood obesity.** *Science* 2006, **312:**279-283.

175. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, et al: **A genome-wide association study identifies novel risk loci for type 2 diabetes.** *Nature* 2007, **445:**881-885.

176. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S, et al: **Genome-wide associations of gene expression variation in humans.** *PLoS Genet* 2005, **1:**e78.

177. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al: **Relative impact of nucleotide and copy number variation on gene expression phenotypes.** *Science* 2007, **315:**848-853.

178. Veenstra-Vanderweele J, Christian SL, Cook EH, Jr.: **Autism as a paradigmatic complex genetic disorder.** *Annu Rev Genomics Hum Genet* 2004, **5:**379-405.

179. Chakrabarti S, Fombonne E: **Pervasive developmental disorders in preschool children: confirmation of high prevalence.** *Am J Psychiatry* 2005, **162:**1133-1141.

180. Ritvo ER, Freeman BJ, Pingree C, Mason-Brothers A, Jorde L, Jenson WR, McMahon WM, Petersen PB, Mo A, Ritvo A: **The UCLA-University of Utah epidemiologic survey of autism: prevalence.** *Am J Psychiatry* 1989, **146:**194-199.

181. Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, Yuzda E, Rutter M: **Autism as a strongly genetic disorder: evidence from a British twin study.** *Psychol Med* 1995, **25:**63-77.

182. **A genomewide screen for autism: strong evidence for linkage to chromosomes 2q, 7q, and 16p.** *Am J Hum Genet* 2001, **69:**570-581.

183. Cantor RM, Kono N, Duvall JA, Alvarez-Retuerto A, Stone JL, Alarcon M, Nelson SF, Geschwind DH: **Replication of autism linkage: fine-mapping peak at 17q21.** *Am J Hum Genet* 2005, **76:**1050-1056.

184. Trikalinos TA, Karvouni A, Zintzaras E, Ylisaukko-oja T, Peltonen L, Jarvela I, Ioannidis JP: **A heterogeneity-based genome search meta-analysis for autism-spectrum disorders.** *Mol Psychiatry* 2006, **11:**29-36.

185. Ylisaukko-oja T, Alarcon M, Cantor RM, Auranen M, Vanhala R, Kempas E, von Wendt L, Jarvela I, Geschwind DH, Peltonen L: **Search for autism loci by combined analysis of Autism Genetic Resource Exchange and Finnish families.** *Ann Neurol* 2006, **59:**145-155.

186. Jamain S, Betancur C, Quach H, Philippe A, Fellous M, Giros B, Gillberg C, Leboyer M, Bourgeron T: **Linkage and association of the glutamate receptor 6 gene with autism.** *Mol Psychiatry* 2002, **7:**302-310.

187. Silverthorn DU: *Human Physiology An Integrated Approach.* San Francisco: Pearson Benjamin Cummins; 2004.

188. Kuemerle B, Zanjani H, Joyner A, Herrup K: **Pattern deformities and cell loss in Engrailed-2 mutant mice suggest two separate patterning events during cerebellar development.** *J Neurosci* 1997, **17:**7881-7889.

189. Powell SK, Rao J, Roque E, Nomizu M, Kuratomi Y, Yamada Y, Kleinman HK: **Neural cell response to multiple novel sites on laminin-1.** *J Neurosci Res* 2000, **61:**302-312.

190. Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, Demeo DL, Murphy A, Su J, Datta S, Rosenow C, et al: **Genomic screening and replication using the same data set in family-based association testing.** *Nat Genet* 2005, **37:**683-691.

191. **Autism Genetic Resource Exchange** [http://www.agre.org/]

192. Reuss B, von Bohlen und Halbach O: **Fibroblast growth factors and their receptors in the central nervous system.** *Cell Tissue Res* 2003, **313:**139-157.

193. Lord C, Rutter M, Le Couteur A: **Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders.** *J Autism Dev Disord* 1994, **24:**659-685.

194. Lord C, Risi S, Lambrecht L, Cook EH, Jr., Leventhal BL, DiLavore PC, Pickles A, Rutter M: **The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism.** *J Autism Dev Disord* 2000, **30:**205-223.

195. Rabbee N, Speed TP: **A genotype calling algorithm for affymetrix SNP arrays.** *Bioinformatics* 2006, **22:**7-12.

196. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: **PLINK: A tool set for whole-genome association and population-based linkage analyses.** *American Journal of Human Genetics* 2007, **81:**559-575.

197. Spielman RS, McGinnis RE, Ewens WJ: **Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).** *Am J Hum Genet* 1993, **52:**506-516.

198. Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM: **Using the noninformative families in family-based association tests: a powerful new testing strategy.** *Am J Hum Genet* 2003, **73:**801-811.

199. Tu IP, Balise RR, Whittemore AS: **Detection of disease genes by use of family data. II. Application to nuclear families.** *Am J Hum Genet* 2000, **66:**1341-1350.

200. Ionita-Laza I, McQueen MB, Laird NM, Lange C: **Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan.** *Am J Hum Genet* 2007, **81:**607-614.

201. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, et al: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296:**2225-2229.

202. Jorde LB: **Linkage disequilibrium and the search for complex disease genes.** *Genome Res* 2000, **10:**1435-1444.

203. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21:**263-265.

204. Duvall JA, Lu A, Cantor RM, Todd RD, Constantino JN, Geschwind DH: **A quantitative trait locus analysis of social responsiveness in multiplex autism families.** *Am J Psychiatry* 2007, **164:**656-662.

205. Alarcon M, Cantor RM, Liu J, Gilliam TC, Geschwind DH: **Evidence for a language quantitative trait locus on chromosome 7q in multiplex autism families.** *Am J Hum Genet* 2002, **70:**60-71.

206. Buxbaum JD, Silverman J, Keddache M, Smith CJ, Hollander E, Ramoz N, Reichert JG: **Linkage analysis for autism in a subset families with obsessive-compulsive behaviors: evidence for an autism susceptibility gene on chromosome 1 and further support for susceptibility genes on chromosome 6 and 19.** *Mol Psychiatry* 2004, **9:**144-150.

207. Bulayeva KB, Glatt SJ, Bulayev OA, Pavlova TA, Tsuang MT: **Genome-wide linkage scan of schizophrenia: a cross-isolate study.** *Genomics* 2007, **89:**167-177.

208. Park N, Juo SH, Cheng R, Liu J, Loth JE, Lilliston B, Nee J, Grunn A, Kanyas K, Lerer B, et al: **Linkage analysis of psychosis in bipolar pedigrees suggests novel putative loci for bipolar disorder and shared susceptibility with schizophrenia.** *Mol Psychiatry* 2004, **9:**1091-1099.

209. Harold D, Jehu L, Turic D, Hollingworth P, Moore P, Summerhayes P, Moskvina V, Foy C, Archer N, Hamilton BA, et al: **Interaction between the ADAM12 and SH3MD1 genes may confer susceptibility to late-onset Alzheimer's disease.** *Am J Med Genet B Neuropsychiatr Genet* 2007, **144B:**448-452.

210. Asai T, Wanaka A, Kato H, Masana Y, Seo M, Tohyama M: **Differential expression of two members of FGF receptor gene family, FGFR-1 and FGFR-2 mRNA, in the adult rat central nervous system.** *Brain Res Mol Brain Res* 1993, **17:**174-178.

211. Courchesne E, Pierce K, Schumann CM, Redcay E, Buckwalter JA, Kennedy DP, Morgan J: **Mapping early brain development in autism.** *Neuron* 2007, **56:**399-413.

212. DiCicco-Bloom E, Lord C, Zwaigenbaum L, Courchesne E, Dager SR, Schmitz C, Schultz RT, Crawley J, Young LJ: **The developmental neurobiology of autism spectrum disorder.** *J Neurosci* 2006, **26:**6897-6906.

213. Kan SH, Elanko N, Johnson D, Cornejo-Roldan L, Cook J, Reich EW, Tomkins S, Verloes A, Twigg SR, Rannan-Eliya S, et al: **Genomic screening of fibroblast growth-factor receptor 2 reveals a wide spectrum of mutations in patients with syndromic craniosynostosis.** *Am J Hum Genet* 2002, **70:**472-486.

214. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, et al: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.** *Nat Biotechnol* 2006, **24:**1151-1161.

215. Boldrick JC, Alizadeh AA, Diehn M, Dudoit S, Liu CL, Belcher CE, Botstein D, Staudt LM, Brown PO, Relman DA: **Stereotyped and specific gene expression programs in human innate immune responses to bacteria.** *Proc Natl Acad Sci U S A* 2002, **99:**972-977.

216. Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C, Wittkowski KM, Piqueras B, Banchereau J, Palucka AK, Chaussabel D: **Gene expression patterns in blood leukocytes discriminate patients with acute infections.** *Blood* 2007, **109:**2066-2077.

Vita

Michael Gormley

Education

Drexel University, Philadelphia, PA
Ph.D. Biomedical Engineering, September 2008 (expected)

Duke University, Durham, NC
B.S.E. Biomedical Engineering May 2004

Research Experience

Computational

Drexel University, Philadelphia, PA
Ph.D. Candidate, 2004-Present

Laboratory

Thomas Jefferson University, Philadelphia, PA
Research Assistant, 2006-2007

Duke University, Durham, NC
Research Assistant, 2003-2004

University of Texas, Austin TX
Research Assistant, 2002 & 2003

Publications

Gormley M, Dampier W, Ertel A, Karacali B, Tozeren A. Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets. *BMC Bioinformatics* 2007, 8:415.

Gormley M, and Tozeren A. Expression profiles of switch-like genes accurately classify tissue and infectious disease phenotypes in model-based classification. *In Press.*