IDENTIFICATION OF GENE EXPRESSION CHANGES IN HUMAN CANCER USING BIOINFORMATIC APPROACHES

by

OBI LEE GRIFFITH

B.Sc., The University of Winnipeg, 2002

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Medical Genetics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2008

Abstract

The human genome contains tens of thousands of gene loci which code for an even greater number of protein and RNA products. The highly complex temporal and spatial expression of these genes makes possible all the biological processes of life. Altered gene expression by mutation or deregulation is fundamental for the development of many human diseases. The ultimate aim of this thesis was to identify gene expression changes relevant to cancer. The advent of genome-wide expression profiling techniques, such as microarrays, has provided powerful new tools to identify such changes and researchers are now faced with an explosion of gene expression data. Processing, comparing and integrating these data present major challenges. I approached these challenges by developing and assessing novel methods for cross-platform analysis of expression data, scalable subspace clustering, and curation of experimental gene regulation data from the published literature. I found that combining results from different expression platforms increases reliability of coexpression predictions. However, I also observed that global correlation between platforms was generally low, and few gene pairs reached reasonable thresholds for high-confidence coexpression. Therefore, I developed a novel subspace clustering algorithm, able to identify coexpressed genes in experimental subsets of very large gene expression datasets. Biological assessment against several metrics indicates that this algorithm performs well. I also developed a novel meta-analysis method to identify consistently reported genes from differential expression studies when raw data are unavailable. This method was applied to thyroid cancer, producing a ranked list of significantly over-represented genes. Tissue microarray analysis of some of these candidates and others identified a number of promising biomarkers for diagnostic and prognostic classification of thyroid cancer. Finally, I present ORegAnno (www.oreganno.org), a resource for the community-driven curation of experimentally verified regulatory sequences. This resource has proven a great success with ~30,000 sequences entered from over 900 publications by ~50 contributing users. These data, methods and resources contribute to our overall understanding of gene regulation, gene expression, and the changes that occur in cancer. Such an understanding should help identify new cancer mechanisms, potential treatment targets, and have significant diagnostic and prognostic implications.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my graduate supervisor Dr. Steven Jones for the opportunity to work at the Genome Sciences Centre and for his generosity, support, encouragement, and guidance during my studies. He was always ready to go to bat for me. I have really appreciated the opportunities to collaborate with excellent researchers and attend numerous conferences to present my work. I would also like to thank Drs. Sam Wiseman and Martin Ester for involving me in such interesting research problems. Thanks also to my thesis committee members Drs. Marco Marra, Angela Brooks-Wilson, Francis Ouellette, and Isabella Tai for their guidance and advice. I am grateful for salary and travel funding from the Canadian Institutes of Health Research, the Michael Smith Foundation for Health Research, the Natural Sciences and Engineering Council, the University of British Columbia (Faculty of Medicine and Department of Medical Genetics), Genome Canada, Genome British Columbia, and the British Columbia Cancer Foundation. I have enjoyed and appreciated the friendship and help of my fellow graduate students, among them Carri-Lyn Mead, Malachi Griffith, Erin Pleasance, Stephen Montgomery, Byron Gao, Monica Sleumer, Anca Petrescu, Simon Chan, Ryan Morin, Angelique Schnerch, Debra Fulton, and Benjamin Good. The research described in this thesis would not have been possible without the help and advice of numerous others at the Genome Sciences Centre and elsewhere including Mikhail Bilenky, Gordon Robertson, Jaswinder Khattra, Sheldon McKay, Gordon Robertson, Peter Ruzanov, Kim Wong, Greg Taylor and everyone in the Open Regulatory Annotation Consortium. I would also like to thank my co-op students, Kristine Pan, Maggie Zhang, Yuliya Prychyna, Bryan Chu, Bridget Bernier and Kathy Kasaian, for all their hard work and stimulating questions. It has been a pleasure to have worked with so many outstanding people. On a personal level, I would like to thank my brothers Malachi and Alex, sister Olivia, Grandma and Grandpa, and the whole Vermette gang for their unconditional support and understanding during this busy phase of my life. Finally, I thank my parents, Rhéa and Ron, and other parent-figures, Werner, Dan, and Veronica, for supporting me, teaching me and allowing me to walk this fascinating path.

Co-Authorship Statement

Together with my supervisor, Steven Jones, I was responsible for the identification and design of the research program described in this thesis. I was primarily responsible for performing the research, data analyses, and manuscript preparation. Portions of the thesis were prepared as part of one or more multi-author publications. For the most part, I have included only the parts of these publications for which I was primarily responsible. However, the co-authors on these publications did contribute analyses, text, figures, tables, editorial suggestions, funding and supervision. Their specific contributions are detailed in the footnotes accompanying each chapter and briefly summarized here. Steven Jones contributed study design, concepts, editorial suggestions, funding and supervision for all chapters. Sam Wiseman contributed text and editorial suggestions to Chapter 1 and concepts, funding, supervision, study design, ethics submissions, data, text and tables to Chapter 4. Adrienne Melck contributed text and editorial suggestions to Chapters 1 and 4. Erin Pleasance, Debra Fulton, Mehrdad Oveisi and Asim Siddiqui contributed concepts, analysis and text to Chapter 2. Martin Ester contributed concepts and provided funding and supervision for Chapters 2 and 3. Byron Gao contributed concepts, code development, analyses, text and figures for Chapter 3. Mikhail Bilenky provided analyses, text and figures for Chapter 3. Yuliya Prychyna contributed background research and text in Chapter 3. Stephen Montgomery contributed concepts, code development, data, analyses and text to Chapter 5. Stein Aerts and Casey Bergman contributed concepts and text to Chapter 5. Many others (e.g., The Open Regulatory Annotation Consortium) made minor contributions to the research described herein (see author lists and acknowledgements in the individual publications for details).

1. Introduction[1,2]

The human genome contains tens of thousands of gene loci which code for an even greater number of different protein and RNA products. The highly complex temporal and spatial expression of these genes makes possible all the biological processes of life from development and differentiation to homeostasis, aging and programmed cell death. Therefore, it is not surprising that altered gene expression by mutation or deregulation is fundamental for the development of many human diseases including cancers. In some cases, cancers can be linked to large changes such as deletions or duplications of entire genes. In other cases, more subtle changes in expression levels of larger numbers of genes are involved. This introduction will review the issues surrounding gene expression analysis and how it can be used to further our understanding of human cancer.

1.1. Thesis overview

The ultimate aim of this thesis was to identify gene expression changes relevant to human cancer. The advent of genome-wide expression profiling technology, such as microarrays, has provided powerful new tools for researchers to identify such changes. As a result, researchers are now faced with an explosion of gene expression profiling data for hundreds of different cancer types and thousands of patient samples. Processing, comparing and integrating these data remains a major challenge. Questions also remain about how and if these data should be combined, especially when produced on different technological platforms. In some cases, computational tools do not scale well to handle these massive datasets. Existing meta-analysis methods assume raw data will be available but this is often not the case. Relating changes in expression to changes in regulation is even more challenging because of our limited understanding of human regulatory control systems. I approached these challenges by developing and assessing novel methods for cross-platform analysis of expression data, scalable subspace clustering, and curation of experimental gene regulation data from the published literature. These methods and resources are generally applicable to any cancer system. I also performed more targeted analyses using tissue microarrays to investigate thyroid cancer. These analyses addressed general questions about the utility of gene expression technologies for identifying

[1] A portion of this chapter has been accepted for publication. Griffith O.L., Melck A., Jones S.J.M., Wiseman S.M. 2007. Thyroid Cancer: Identification of Gene Expression Markers for Diagnosis. In: Hayat MA, ed, *Methods of Cancer Diagnosis, Therapy and Prognosis*. Springer Publishing Company. New York, NY.

[2] Co-authorship details: I was primarily responsible for writing all text in this chapter. However, contributions and editorial suggestions were made throughout by Sam Wiseman, Adrienne Melck, and Steven Jones.

biologically important relationships through comparisons between technologies; validation of expression data against independent knowledge sources such as the Gene Ontology; and translation of significant findings at the RNA level to the protein level.

1.2. Gene expression and gene regulation

For the purposes of this thesis, 'gene expression' refers to the process by which information encoded in the genome (e.g., the DNA coding for a gene) is converted into the functional molecules present and active in a cell (usually a protein, but sometimes RNA). 'Gene regulation' on the other hand refers to all the mechanisms or processes that control the rate, location or manner in which genes are expressed. With the human genome sequence nearly completed [1] and many of its genes identified, biologists are turning to the search for the genome's second code – the gene regulation code [2]. It is increasingly likely that the evolution of genetic diversity in animal forms is as much the result of differences in gene regulation as in the genes themselves [3]. To illustrate, compare the gene regulation systems of yeast and mammals. The yeast genome contains approximately 300 transcription factors. These sequence-specific DNA binding proteins usually bind regulatory sequences immediately 5' of the transcription start site. A typical example would include a TATA element which serves as a binding site for a TATA-binding protein (TBP). The binding of TBP in turn is regulated by two or three upstream activating sequences (UAS) for one or two additional sequence-specific transcription factors (TFs). In most cases, the entire regulatory region is contained within a few hundred bases of the gene. In contrast, a metazoan genome (like the human genome) contains as many as 3,000 transcription factors and highly complex sets of regulatory elements (Figure 1.1; also described further below). Despite the obvious importance of gene regulation in biological systems, our ability to identify and predict functions of regulatory elements remains limited [4]. Thus, few mutations in these sequences have been linked to disease susceptibility or progression despite their potential role in many human disorders.

1.2.1. Levels of gene expression and gene regulation

Messenger RNA (mRNA) is the molecular intermediate between DNA and protein. As such, the quantity of mRNA is often used as a measure of a gene's expression level and assumed to be related to the amount of functional end product (e.g., protein) for that gene. However, the pathway from gene to functional protein in humans is a complex process involving several stages including: chromatin decompaction, initiation, elongation, termination, transcript cleavage, 5'

capping, 3' polyadenylation, splicing, mRNA packaging and export (from nucleus to cytoplasm), translation, protein folding, and post-translational modification (Figure 1.2) [5]. While often presented as independent steps, in reality, each stage is part of a continuous process, physically and functionally interconnected. Some stages are regulated by protein (or RNA) factors which in turn are regulated by other factors, extracellular signalling molecules, or environmental cues. Others are regulated by epigenetic modifications such as DNA and histone methylation and acetylation [6, 7]. Much of the regulation involves binding of regulatory factors to DNA regulatory sequences. A typical gene might contain several enhancer sequences, 5', 3' or intronic to the gene. Each enhancer could span ~500bp and contain many binding sites for multiple different transcription factors (activators and repressors). The core promoter might contain 3 or more sequence elements to recruit the transcription complex. Other proximal promoter elements might act as tethering agents to recruit distal enhancers. And, finally, silencer elements prevent enhancers associated with one gene from inappropriately regulating neighbouring genes. In mammals, all these elements can be spread over distances of 100kb or more. This complex organization allows detailed control of gene expression to meet the environmental, temporal, cell, and tissue-specific needs of the organism [8]. Given the level of regulatory complexity outlined above, it is important to question whether a simple measurement of RNA levels can be a meaningful measure of final protein activity. The emergence of high-throughput protein detection methods has recently allowed more global comparisons of protein and mRNA levels. The Pearson correlation coefficients for these comparisons range from 0.46 to 0.76 but still represent relatively small numbers of genes/proteins (70 to 678) [9]. These numbers will likely improve as proteome-wide protein detection methods are further developed. Until then, we will remain dependent on mRNA detection as a proxy for protein activity in genome-wide studies. While clearly not perfect, this introduction will highlight a number of studies that have demonstrated real and practical successes with genome-wide transcript-level profiling.

1.3. Gene regulation analysis

Identification of coregulated genes and their regulatory elements has been the subject of numerous studies. The approaches for identifying regulatory elements can be broadly grouped into computational and experimental methods. The computational identification of regulatory networks is a complicated and error-prone process but often considered necessary due to the sheer size and combinatorial nature of the problem. There are numerous computational approaches for the identification of transcription factor binding sites. Surveys of relevant web-

based tools are available from several investigators [10, 11] and a comparative evaluation of many of the popular algorithms was reported [12]. Some of the general approaches will briefly be reviewed here. Pattern matching (motif scanning) can be used when sufficient prior knowledge exists for a protein's binding site preferences. This is usually in the form of a matrix or consensus pattern representing the previously observed nucleotides at each position for the site. Such binding models are derived from experimental assays and can be obtained from databases such as Jaspar [13] or TRANSFAC [14]. The genome of interest is then scanned for sequences with similarity to the binding model represented by the matrix/consensus. High scoring matches are considered potential binding sites for that protein. Other algorithmic approaches have been developed for de novo prediction (i.e., identifying new motifs without using a known binding model). These methods typically involve some kind of recurrence or over-representation analysis that depends on "real" sequences being conspicuously common compared to non-functional sequence. A variety of heuristics are employed such as Hidden Markov Models, Gibbs sampling, expectation-maximization, or even exhaustive enumeration [10]. Other popular approaches make use of sets of sequences from orthologous ('phylogenetic footprinting') or coregulated genes (as determined by coexpression for example). Sites that are evolutionarily conserved or consistent between tightly coexpressed genes are hypothesized to be functional. The methods described above have been used in numerous variations and combinations with some success. However, all suffer from a limited amount of training data, often result in high rates of false-positive predictions, and must be experimentally validated before they can be trusted with high confidence.

Numerous experimental methods have been developed and widely used over the last few decades to dissect promoter regions, identify transcription factor binding sites, and characterize protein-DNA binding interactions. These methods include protocols such as DNaseI hypersensitivity assays, reporter gene assays, electrophoretic mobility shift assays (EMSA), and chromatin immunoprecipitation to name just a few of the most popular [10]. Some methods measure indirect clues to transcription factor binding such as chromatin status (e.g., DNaseI hypersensitivity) or measure the effect of sequence changes on expression level (e.g., reporter gene assays) without knowing the exact regulatory proteins involved. Others measure the protein/DNA interaction without necessarily pinpointing the critical nucleotides involved. In many cases, combinations of several assays under numerous conditions or cell types are needed to really understand the complexity of temporal- and tissue-specific regulatory interactions. Of

course, this is time consuming and not easily scaled to the vast number of factors and sequences thought to be functional in human cells. Furthermore, many of the findings of such studies are not computationally accessible and therefore not amenable to the more global analysis necessary for a unified understanding of gene regulation.

Several high-throughput methods for investigating interactions between proteins and DNA *in vivo* have recently emerged that may solve many of the problems faced by computational and low-throughput experimental approaches. These methods, sometimes referred to as genome-wide location analyses, are performed using a variety of protocols, including ChIP-chip, ChIP-PET, and ChIP-seq [15-17]. In all three cases the first step is a chromatin immunoprecipitation (ChIP). DNA-binding proteins are first cross-linked to DNA (e.g., with formaldehyde). Then, chromatin is isolated and the DNA sheared along with bound proteins into small fragments. Antibodies specific to the DNA-binding protein are used to isolate the complex by precipitation. Next, the cross-linking is reversed to release the DNA and digest the proteins. Finally, the DNA is often purified and amplified using PCR. At this point, the three protocols diverge in terms of their sequence identification methods. In ChIP-chip, the oldest of the methods, the purified DNA and appropriate controls are fluorescently labelled and applied to arrays of DNA probes (microscope slides) for microarray analysis. Tiling microarrays are populated with probes spanning entire chromosomes or genomes. Strong hybridization of labelled DNA to specific probes identifies the bound sequences. In ChIP-PET and ChIP-seq, libraries are constructed for DNA sequencing of the bound sequences which can then be mapped to their genomic locations by alignment. ChIP-PET libraries have typically been sequenced with standard Sanger sequencing and produce short paired-end-tag (PET) sequences which correspond to the 5' and 3' ends of the bound region. ChIP-seq libraries on the other hand are typically sequenced with the so-called 'next generation' sequencing technologies producing vastly increased numbers of short sequence tags. Instead of PETs, 'peaks' of overlapping sequence fragments identify the bound region. However, as ChIP-PET groups migrate to the newer sequence technologies and ChIP-seq groups adapt their protocols for PETs to improve alignment success, these two methods have begun to fuse into a single approach [18]. It is hoped that improvements in regulatory element detection together with gene expression analysis will someday lead to a comprehensive understanding of gene expression and its regulation.

1.4. Gene expression analysis

1.4.1. Gene expression technologies

A number of technologies currently exist for large-scale profiling of gene expression at the level of transcription. The most common platforms fall into two categories: spotted cDNA microarrays [19] and oligonucleotide arrays [20]. In addition to microarrays a number of other high-throughput methods exist such as Serial Analysis of Gene Expression (SAGE) [21] and massively parallel signature sequencing (MPSS) [22]. Each platform is capable of quantifying the transcript levels of hundreds to tens of thousands of genes simultaneously. In the past 10 years approximately 20,000 papers and hundreds of reviews related to microarrays have been published covering every aspect of their use from construction to data analysis and their application in cancer [23-31]. The two major microarray technologies, along with SAGE and new tag-sequencing approaches will briefly be presented here.

1.4.1.1. cDNA microarrays

Modern cDNA microarrays are constructed by PCR amplification of cDNA or genomic clones and spotting of the amplification products onto a solid support using robotic printers. Microarrays with thousands of spots can be generated by this method. Sample detection involves hybridizing the target mRNA (most commonly as cDNA after reverse transcription), labelled with fluorescent or radioactive nucleotides, to the probe DNA on the array. The most common protocol involves differentially labelling control and test samples with two fluorescent dyes such as Cyanine 3 (Cy3) and Cyanine 5 (Cy5). Lasers of the correct wavelengths (550nm for Cy3; 649nm for Cy5) will cause fluourescent excitation, giving off light at specific wavelengths (570nm for Cy3; 670nm for Cy5), that can be detected by a scanning microscope and quantified by the scanner software. By calculating the ratio of Cy3 to Cy5 signal (or vice versa), the mRNA abundance of the test sample relative to control sample can be determined.

The initial advantage of spotted cDNA arrays was their flexibility. Custom arrays with any arbitrary gene set could be constructed as long as clones were available. Many labs constructed their own arrays. This allowed construction of highly specialized arrays targeted at specific cancers or pathways. As annotation of the human genome progressed and sequence-verified clones became available, whole genome expression profiling became possible. The difficulties of clone management and array production on a genome-wide scale were also solved when commercial solutions became available. However, some of these commercial products were

disadvantaged by not providing the sequence data for each clone and having poor annotation of clone sequences. Technical disadvantages arose from the presence of repetitive elements and cross-hybridization between homologous gene families. Except in special cases, cDNA arrays have been largely supplanted by oligonucleotide arrays which have several important advantages.

1.4.1.2. Oligonucleotide arrays

Oligonucleotide (oligo) arrays can be constructed by numerous methods including spotting pre-synthesized oligos, *in situ* oligo synthesis by photolithography, or maskless array synthesis on a glass slide [32]. There are a large number of commercially available oligo arrays from manufacturers which include: Affymetrix, Illumina, Nimblegen, and others, each with their own advantages and disadvantages. Development of the above methods has allowed extensive miniaturization of array construction making it possible to construct arrays with hundreds of thousands of spots. For example, the current version of Affymetrix GeneChips (Human Genome U133 Plus 2.0) includes over 1 million distinct oligonucleotides representing more than 47,000 human transcripts. Unlike cDNA arrays, many oligo arrays produce an intensity signal that allows absolute (as opposed to relative) quantification. Intensity is measured by laser scanning and fluorescence excitation, similar to cDNA arrays, but typically with only one sample and one type of dye hybridized per array. The ability to custom design large numbers of oligonucleotide sequences in parallel makes it possible to avoid repetitive sequences and create unique probes even for highly homologous gene families. For these reasons, oligonucleotide arrays have become the preferred microarray type. The use of standardized platforms such as the Affymetrix GeneChips has created new opportunities for inter-laboratory comparisons and meta-analyses whereas flexible and customizable arrays such as the Nimblegen products allow more frequent sequence updates, advanced, cost-effective and custom designs, and rapid application to recently sequenced genomes.

1.4.1.3. SAGE

Serial analysis of gene expression (SAGE) is a sequencing-based approach to gene expression profiling [21]. Briefly, SAGE involves the extraction of a short sequence (tag) from each polyadenylated (polyA) RNA. This is accomplished by conversion of polyA RNA to cDNA followed by a series of restriction digestions and ligations to produce double stranded ditags (pairs of tags) with PCR linkers. The ditags are PCR amplified and the linkers released. Next, the

ditags are concatenated, cloned and finally sequenced. The resulting sequences are mapped by alignment to the transcriptome and the frequency of any particular tag sequence is taken as a quantitative measure of the source transcript's abundance. There are several data analysis issues and potential sources of bias in the SAGE method such as sequence biases in library construction, PCR amplification bias, production of non-canonical tags (unexpected restriction sites), sequence errors, variation in tag length and mapping ambiguity [33]. Also, due to the high cost of sequencing, SAGE libraries are sometimes not sequenced to depths sufficient for robust detection of low-abundance transcripts. Pleasance and Jones (2005) recently reviewed the use of SAGE for transcriptome studies [33]. SAGE can be quite reproducible, is not prone to problems such as cross-hybridization, produces absolute estimates of expression abundance, and can have detection efficiency similar to oligo arrays with sufficient sampling depth. Also, SAGE does not require *a priori* knowledge of the genes to be profiled. Thus, in addition to profiling known genes, SAGE libraries are a valuable resource for the identification of novel genes and transcripts. However, costs can be considerably greater than microarray profiling. Thus, the number of replicates is generally small (often no replicates are done). This together with the sample depth issue can make statistical analysis of SAGE data challenging.

1.4.1.4. Next-generation tag-sequencing methods

The recent development of 'next-generation' sequencing technologies promises to build on the advantages of tag sequencing based approaches like SAGE and address the issues of sampling depth and cost. An often cited target for these developments is the '$1,000 genome', a fully sequenced human genome for the price of only $1,000. This would require a 10,000-fold reduction in cost over conventional Sanger sequencing methods [34]. Current developments such as the Solexa Illumina system (www.solexa.com), 454 system (www.454.com) and the polony sequencing method from George Church's laboratory [35] are still under development but do promise approximately 100-fold cost reductions [34]. In these systems, single DNA molecules (usually pre-fragmented) are clonally amplified in spatially separate locations on a highly parallel array and used as templates for sequencing by synthesis. The three systems use different sequence chemistry and have different potential for scalability but all produce vast quantities of short high-quality sequences. Thus, application of these technologies to analysis of RNA samples has the potential to create a SAGE-like library of short sequence tags at much greater sampling depth and a fraction of the cost of SAGE. Ng *et al.* (2006) demonstrated this potential using the 454 platform on a paired-end tag (PET) library constructed from MCF7 cells [18]. To validate

the method's quantitative ability they performed qRT-PCR on a selection of 12 PETs with a wide abundance range and showed good correlation. Complete sequencing of the transcriptome using next generation sequencing, has the potential to provide a more quantitative and sensitive enumeration of transcript abundance whilst also identifying spliced variants and polymorphisms [36]. In the near term these new sequencing-based approaches will likely be of greater interest for basic researchers, whereas hybridization-based methods are more likely to serve a clinical role for reasons of speed and simplicity. However, in the long term, sequencing-based methods may replace microarrays in the clinic.

1.4.2. Experimental issues

1.4.2.1. Array design issues

There are many different experimental design issues to consider when selecting or designing an expression profiling experiment. Two-color microarray experiments can be carried out as a direct comparison, balanced block design, reference design, or loop design (as reviewed by Quackenbush (2005) [37]). In many cases it is advisable to conduct 'dye swaps' where the same two samples are hybridized to the array with opposite labelling. This eliminates 'dye effects' where one dye is consistently detected at higher intensity levels irrespective of the actual abundance. Single-color microarray experiments have only one sample per array but there still may be a large number of possible comparisons between arrays to consider such as different stages or disease pathologies. Some arrays are designed with positive and/or negative hybridization controls. The positive controls are typically genes thought to have relatively constant expression (i.e., 'house-keeping' genes). Negative controls might be random sequences or genes known to not be expressed (e.g., from a completely different species). These control probes serve as a quality control to identify problems such as defective arrays, bad samples, or poor hybridizations. They can also provide a useful reference for normalization. Another kind of control to consider is a 'spike-in'. By adding a known quantity of a particular transcript to every sample, hybridization conditions can be assessed and potential biases identified [38]. A major design issue to consider is hybridization bias. Microarrays rely on the assumption that signal intensity is directly related to amount of hybridization which is in turn directly related to amount of transcript complementary to the probe sequence. In reality, hybridization is a complex process with overall binding affinity affected by the length, nucleotide composition, labelling protocol, and secondary structure characteristics of the probe sequence. These issues can be minimized but not eliminated by using variable length probes carefully selected from the transcript sequence to

more closely match the binding affinity across the array and avoid unfavourable secondary structures. Some custom oligo arrays are able to incorporate this in their microarray designs.

1.4.2.2. Sample preparation, collection and storage

Once the experimental design is finalized, high quality samples are needed. A large number of variables can affect sample preparation including the media used for growing cells, RNA preparation method, amount of source tissue, sample handling, etc. Sample quality and quantity should be tested after RNA extraction or amplification from source sample, cDNA generation (the amount of cDNA should be close to the amount of starting RNA), and label incorporation. Well established protocols and instrumentation exist for this purpose.

1.4.2.3. Replicates and reproducibility

Ultimately, an unknown amount of variability will exist in the expression patterns obtained from different samples. It is important to assess the level of this variability both technically and biologically. However, with modern array technologies the level of technical variation has been significantly reduced. As technical variation is reduced, biological replicates become more useful and technical replicates can be minimized. Sufficient replicates are also necessary to accurately determine the level of background intensity. Only by assessing replicate variability are meaningful levels of statistical confidence assigned to the differences or patterns observed. There is no single standard number of replicates for an experiment. A balance must be found between limiting factors such as the cost of experiments, availability of samples, and the power to detect statistically significant events. Generally, for a direct comparison between two conditions, at least three replicates are needed for reproducible identification of differentially expressed genes [39]. However, statisticians commonly recommend considerably more than three replicate experiments. A simple rule to follow is to ensure at least five degrees of freedom (df) where df is the number of independent units minus the number of distinct treatments in the experiment [40]. For example, in a comparison of four tumour tissue samples to four normal tissue samples there are six dfs. The eight tissue samples represent eight independent experimental units and the two tissue types (normal and tumour) represent two experimental conditions (8 - 2 = 6 dfs).

1.4.3. Data analysis issues

The successful implementation of expression profiling analysis requires consideration of not only the various laboratory protocols employed, but also the computational issues of data

collection, processing, storage and statistical analysis. A typical microarray study can produce over a million data points (e.g. 20 samples on HGU133 Plus 2.0 array with 54,000 probe sets). The development of computational methods to deal with these massive datasets is an area of active research with a large number of databases, software packages, and algorithms now available for gene expression analysis. It is not yet clear if any standard protocols for data analysis will emerge. In some cases, the application of several different algorithms allows for the exploration of different aspects of the data. A review of some of the commonly used approaches for microarray data analysis has recently been reported [40].

1.4.3.1. Quality assessment

Before proceeding to any further analysis, it is customary to perform some preliminary quality assessment of the microarray experiment. This allows the filtering out of problem arrays or spots. For example, a researcher may wish to discard arrays with a high percentage of missing values or discard probes that are missing in a large proportion of samples. Image plots can be used for visual inspection to identify spatial irregularities or imperfections (e.g., fingerprints, or flaws on the array). Box plots express the distribution of intensities and the number of outliers. Ratio-intensity plots or MA-plots can be used to assess systematic bias on intensity values [37]. It is a good practice to exclude outliers before normalization as most normalization methods assume that gene expression is relatively constant across samples. This is not always true and should be evaluated.

1.4.3.2. Normalization and background correction

After basic image processing, normalization and background correction are generally the first steps in preparing expression data for analysis. The purpose of normalization is to remove sources of technical variation by adjusting for differences in labelling and detection efficiencies and for differences in the quantity of RNA hybridized to different arrays. Background correction attempts to eliminate the effect of low-level noise inherent to the array and produce an accurate estimate of the actual transcript expression level. Many normalization methods include a background correction or summarization method. Most normalization methods can be applied either globally (to the entire dataset) or locally (to some subset of the data such as each subgrid of the array). Three commonly used normalization strategies for two-color arrays are: (1) total intensity normalization, (2) normalization using regression techniques such as LOWESS, and (3) normalization using ratio statistics (reviewed in [40]). For single-channel arrays, quantile

normalization or some variation is most commonly utilized. Methods such as GCRMA start with quantile normalization but also model base-specific effects such as the stronger bonding of G/C pairs [41]. A detailed review of normalization strategies for both one- and two-channel arrays is available [42]. To compensate for different library sizes, SAGE tag counts are commonly normalized to 10,000 or 1,000,000 tags/library as follows: Tag frequency = (tag count x 1,000,000)/total tags in library [43]. After normalization, the data for each gene are commonly expressed as the logarithm of the normalized value (ratio, intensity or tag frequency). This makes the variation of intensities less dependent on the absolute magnitude and evens out highly skewed distributions. It has been suggested that the choice of normalization has relatively small effects on the final analysis outcome of larger microarray studies (compared to platform choice, RNA quality, etc) but can have important effects for smaller studies [44].

1.4.3.3. Probe/tag mapping

A critical preliminary step to expression profiling analysis is the accurate mapping of probe, clone, or tag sequence to the correct gene locus or transcript. Mapping errors are common, can lead to misidentification in differential expression studies and are a frequent source of discrepancy between platforms [45, 46]. With genome annotations constantly being revised, it is important to periodically update the mapping information for any platform. Some commercial platforms such as Affymetrix periodically release updated annotation files. However, there are also a number of public resources such as SOURCE[47], DAVID [48], DRAGON [49], and RESOURCERER [50] for array probes and SAGEmap [51], and DiscoverySpace [52] for SAGE tags. Another difficulty is that there are several target gene identifiers that a researcher may wish to utilize (Entrez, Unigene, Refseq, Uniprot, etc). Often, a probe sequence will be mapped by its sequence to an intermediary identifier and then cross-reference tables used to map (sometimes through several steps) to the final target identifier. This introduces increasing possibilities for error as cross-reference tables may not be current. Ideally, direct mapping by sequence from the probe to the desired target identifier should always be used as this will guarantee the most current mappings and best correlation with other datasets or platforms [53].

1.4.3.4. Differential expression analysis

Expression profiling experiments are most commonly applied to the problem of identifying genes that are expressed differently between two distinct conditions (e.g. cancer versus normal tissue). This typically involves categorizing samples into two groups, determining some measure

of difference between the two categories for each gene (e.g. a mean fold-change), assigning a measure of statistical significance to each gene, and determining a cut-off value to select a final list of "interesting genes".

1.4.3.4.1. Types of comparisons

The comparison of two conditions (e.g. cancer versus normal) is the most straightforward case. In two-condition cases, the conditions can be either independent (e.g. two different patient populations, one cancer and one normal) or dependent (e.g. a set of matched normal and cancer samples from the same patients). In more complex comparisons there may be multiple conditions. These can also be categorized as independent (e.g. four patient populations, one for each of four cancer stages) or dependent (e.g. a related set of samples taken from patients at different time points). Each of these situations requires different statistical considerations.

1.4.3.4.2. Statistical considerations

In cases where no replicates are available, comparisons are limited. The fold-change can be calculated between the two conditions and a fixed cut-off chosen with the most common threshold being two-fold change. In SAGE analysis, where a lack of replicates is common, the Audic-Claverie statistic is often used [54]. This has the advantage of considering the respective library sizes and tag counts to assign a p-value to the observed tag-count difference. If replicates are available, many more options become available. A number of different statistical tests are available to assess differential expression in two-condition comparisons such as a simple t-test or a non-parametric equivalent such as a Wilcoxon rank-sum test. In cases where the samples are related to each other a paired t-test or Wilcoxon matched pairs signed-rank test could be used instead. For multi-condition comparisons more sophisticated statistics such as ANOVA are employed. A common problem with these statistics arises when measures of RNA abundance have small intensity differences and extremely low variance. These genes are not likely of interest but tend to have very significant p-values because of their low variance. A number of variations of the t-statistic have been proposed that use different penalizing factors to overcome this problem. Perhaps the most popular solution is the 'significance analysis of microarrays' (SAM) method [55]. SAM is now able to deal with all four basic situations (two- or multi-condition, independent or dependent). Using several different approaches may also be a good strategy. No single approach is likely to be completely correct but if a gene passes multiple tests

results are perhaps more likely to be 'real'. Reviews of the statistical approaches to differential expression analysis for microarrays [56] and SAGE [57] have been published.

1.4.3.4.3. Multiple testing issues

Multiple testing is a major concern for modern genome-wide expression profiling. Typically, researchers choose a P-value cut-off of 0.05 and assume that all genes with a lower P-value will be of interest. But, supposing that the array contains 10,000 genes to test, we can expect that approximately 500 (5% of 10,000) of these genes would show a P-value of less than 0.05 by chance alone. Therefore, some type of multiple testing correction or test reduction (or both) should always be carried out. Test reduction can be accomplished by pre-filtering the gene list based on some basic criteria, for example, genes with very low variance across conditions. Multiple testing correction involves correcting P-values so that a given false positive rate (type I error rate) is guaranteed for all tests. Methods attempt to either control the family wise error rate (FWER: probability of at least one type I error) or the false discovery rate (FDR: expected proportion of type I errors in the rejected hypotheses). A commonly used, but extremely conservative, approach is the Bonferroni correction where the P-values are simply divided by the number of tests performed. This method is so conservative that it is not uncommon to lose all significant results after its application. For this reason, less stringent procedures such as the Benjamini and Hochberg method [58] and others have also become popular (for review see [59]).

1.4.3.5. Clustering analysis

After differential expression analysis (described in section 1.4.3.4), the most common use for expression profiling data is probably clustering analysis. Essentially, this involves identifying groups of genes that appear to be 'coexpressed' by grouping genes with similar gene expression patterns. First, a similarity measure is calculated between all gene pairs and then genes are grouped by clustering or partitioning based on those similarity measurements. Dozens of different similarity measurements and clustering algorithms exist and there is often disagreement as to which is best [60]. Perhaps the most common combination used is the Pearson correlation coefficient as a similarity measure and some form of hierarchical clustering as the clustering algorithm. The Pearson correlation is a measure of correlation between two variables commonly used across many research fields. An advantage of the Pearson correlation is that it is not sensitive to scaling or differences in average expression level. Another commonly used similarity

measure is the Euclidian distance. Hierarchical clustering can start with a single cluster that is progressively subdivided into smaller clusters (divisive clustering) or it can start with single-gene clusters and successively join the closest clusters until all genes have been added to a single super-cluster (agglomerative clustering). Figure 1.3 shows a cartoon depiction of how raw data might be clustered using an agglomerative hierarchical clustering method. A whole family of clustering methods differs only in how they determine the inter-cluster distance for joining or subdividing. For example, single linkage defines the distance between clusters as the shortest distance between any two cluster members. In all cases, the final product is a tree-shaped data structure, or dendrogram in which each element and sub-cluster is connected to all others as part of one or more superclusters. In order to obtain discrete clusters, the tree must be cut at a given level. Partitioning methods on the other hand (such as k-means clustering) subdivide the data into a pre-determined number of subsets without any implied hierarchical relationship between the clusters. Determining the correct number of clusters can be a difficult and arbitrary task. Figure 1.4 illustrates a simple case where cluster analysis is applied to gene expression data. First, the genes are clustered using hierarchical clustering. Then, the samples are partitioned using k-means clustering (with k=2). Yet another class of clustering methods termed biclustering or subspace clustering attempts to cluster both genes and experiments/conditions simultaneously. These will be discussed further in chapter 3. For an excellent review of clustering methods for microarray gene expression data, see [61].

Genome wide clustering analyses in *C. elegans* and *S. cerevisiae* have been used with some success to identify genes that are coregulated or share a common function [62-64]. In human systems however, this "guilt-by-association" approach has received criticism because of high levels of noise and other problems inherent to the methods [65]. Thus, methods of filtering, assessing confidence, and refining clustering predictions are needed. An approach that has gained recent popularity is to use multiple sources of biological information together with expression data to define more biologically meaningful modules of coregulated genes [66-68]. This raises important questions about how different datasets should be combined and how clusters can be assessed as biologically meaningful. Clustering has also been used in cancer studies to group patients according to their expression profiles to aid in developing a molecular classifier or to define to new molecular subtypes for the cancer (discussed further in sections 1.5.2 and 1.5.3).

1.4.3.6. Classification analysis

Classification and clustering are closely related topics. However, they do have distinctly different methods and purposes. Clustering is generally an unsupervised process where genes or samples are grouped according to their similarities in expression patterns. Classification on the other hand is a supervised process that seeks to identify specific gene expression patterns which can discriminate one set of samples from another (e.g. cancer versus normal patients). Classification in cancer will be specifically addressed further in section 1.5.2. For any kind of classification problem, numerous statistics and algorithms exist. Even a t-test for two classes can be used (as in differential expression analysis) to identify a gene with discriminatory potential between two classes. In the simplest case, a gene that is found to be expressed in one class but not in the other can be used as a single-gene classifier. However, the situation is rarely this simple and classification algorithms generally make use of multiple genes and more complex rules to distinguish between classes. Some popular classification algorithms for cancer analysis include support vector machines (SVM), receiver operator characteristic (ROC) regression, and tree or forest based methods [69].

Once an algorithm has been selected, it is generally 'trained' on the available samples to optimize its discriminatory power. Ideally, the classifier would then be validated by application to an independent test set (i.e., samples not used in training). However, sufficient samples are often not available for both training and independent testing [70]. In such cases, a cross-validation method (e.g., leave-one-out cross-validation) is used to assess performance. A small subset (k) of samples are left out of the training set, the algorithm is trained on the remaining samples (n-k) and the resulting classifier applied back to the k samples as a test set. This entire process is then repeated for multiple different random divisions into k and n-k subsets to produce average measures of performance.

Classification performance is typically reported using sensitivity, specificity, accuracy (or error rate) and ROC curves. Table 1 shows a simple example to explain some of these concepts. Sensitivity refers to the percentage of patients who are classified as having the disease among the group of patients who actually do have the disease. Specificity refers to the percentage of patients who are correctly classified as being disease-free among the group of patients who do not have the disease. Accuracy refers to the percentage of correct classifications (true positives

and true negatives) among all patients. Ideally, sensitivity, specificity, and accuracy would all be 100%. However, this is rarely the case and therefore, the relative importance of sensitivity and specificity must sometimes be balanced depending on the nature of the disease. In cases where untreated disease is associated with high mortality (e.g., certain aggressive cancers) even 99% sensitivity might be unacceptable because the 1% of patients sent home as false negatives might die without treatment. On the other hand, for diseases with high treatment costs (e.g., cancers requiring a hysterectomy) there might be a greater emphasis on specificity in order to minimize the damage of false positives. In cancer, both untreated disease and unnecessary treatment can bear high burdens. Unfortunately, with many classifiers, sensitivity and specificity are dependent on the specific thresholds used to convert a continuous score into a binary result (i.e., 'disease' or 'no disease'). Typically, by increasing the threshold, sensitivity can be improved at the expense of specificity or vice versa. A common technique for choosing optimal thresholds and generally evaluating the utility of the classifier across a range of thresholds is to plot a ROC curve. The ROC curve is created by plotting sensitivity versus 1-specificity for a continuous range of classification thresholds. Different methods are then used to choose the best balance between sensitivity and specificity. The area under the ROC curve (AUC) or ROC integral is often considered a good measure of the overall performance of the classifier and can be used to compare performance of different classifiers.

1.4.3.7. Expression analysis software and databases

There has been an explosion of algorithmic implementations, software packages, and databases for the analysis of gene expression data in the bioinformatics field. Here, I will discuss just a few of the most common analysis tools based on my own experience. For expression profiling analysis from array processing to normalization, differential expression analysis, clustering and classification I have utilized the R programming language (http://www.r-project.org) and the associated Bioconductor packages (http://www.bioconductor.org). This is a free, open-source resource with a steep learning curve, but extremely versatile and powerful. Some useful packages are 'gcrma' for background correction and normalization of Affymetrix data, 'limma' for differential expression analysis of both one- and two-color arrays, 'samr' for SAM analysis and 'multtest' for multiple testing correction methods. A large number of similarity measures, clustering algorithms, and classifiers are also available as standard functions in R. For very large datasets (especially for SAGE data which can have many tags) the open source clustering library [71] is useful because of its speed and memory efficiency. A large number of classification

algorithms are also available in the Weka package [72, 73]. For software packages with a graphical user interface (GUI), the TM4 suite [74] for cDNA microarray data and Dchip [75] for oligonucleotide array data are also available. A good commercial solution for classification with a GUI is the Random Forests software from Salford Systems (San Diego, CA, USA). Random Forests classification and ROC curve generation can also be performed using the R libraries 'randomForest' and 'ROCR'.

In addition to software for analyzing expression data, a number of databases exist for the submission, storage, and dissemination of raw expression data for the research community. The Gene Expression Omnibus (GEO) currently holds more than 147,000 samples for more than 100 different organisms [76], the Stanford Microarray Database (SMD) contains more than 10,000 public experiments for more than 20 organisms [77], and ArrayExpress contains more than 2,000 experiments [78]. Each is capable of housing multiple different platform types. Another useful database is Oncomine [79], a database specializing in cancer expression datasets and web-based analysis tools that contains more than 20,000 microarrays representing 39 different cancer types.

1.4.3.8. Open-access, open-source, standards and ontologies

An important development in bioinformatics and arguably all of life sciences is the movement toward open-access publications and data, open-source software code, data reporting standards and ontologies. Most work in bioinformatics is dependent on freely available gene expression or other kinds of high-throughput genomic data. Even low-throughput experiments would ideally be made available freely through scientific literature (without subscription) in order to facilitate curation into centralized repositories or text-mining analyses. Such efforts are also dependent on increased standardization and formalization in how information is reported. For example, standards like the 'minimum information about a microarray experiment' (MIAME) ensures that the results from a microarray study can be easily interpreted and verified [80]. MIAME consists of detailed descriptions of experimental design, array design, samples, hybridizations, measurements, and normalization controls. This standard has gone a long way to ensuring that microarray results made publicly available will actually be useful to the public. Another important development has been the development of controlled vocabularies and ontologies. A controlled vocabulary refers to a limited set of standardized terms for describing a system. An ontology (which often incorporates a controlled vocabulary) is an explicit representation of such terms. For example, the Gene Ontology (GO) defines relationships between standardized terms

to describe the biological process, molecular function, and cellular localization of gene products [81]. Other ontologies such as EVOC [82] and SNOMED [83] describe cell types and patient records respectively. Whereas GO has almost completely revolutionized gene expression analysis, much work is still needed to further develop this ontology and others. In particular, ontologies for describing experimental protocols and tissue samples are needed.

1.4.4. Validation methods

While expression profiling technologies such as microarrays are well recognized as powerful tools for identifying expression changes, they are also known to generate a high number of false positives. Therefore, it is common practice to validate microarray findings with an independent assay. There are two major levels at which validation is desired. First, there is technical validation of the actual microarray measurement (i.e., can the changes in transcript level be confirmed?). Secondly, there is validation at the biological level (i.e., do the changes have something to do with the condition being studied?). In the first case, the original samples might undergo another comparison using a more precise quantitative assay. In the second case, validation might occur on an independent set of samples to show that the gene effect is generalizable or a functional assay could be designed to test the hypothesized effect of the gene expression change (e.g., a gene knockout in a model system for the disease). Perhaps the most common technical validation method is (semi-)quantitative real time PCR (qrtPCR) which permits for validation at the RNA level. For an extensive review of qrtPCR and its application to cancer see [84]. Other RNA validation techniques include northern blot analysis and in situ hybridization.

Another approach for validation is at the bioinformatic level. A common analysis is to evaluate a list of differentially expressed genes for over-representation of specific biological processes or pathways using the Gene Ontology or KEGG resources. The significant observations can then be assessed in the context of the current literature for the disease being studied in order to determine if expected pathways or processes are activated or deactivated. Other bioinformatic validations involve searching for concordance or significant overlap with previously published datasets or other data types (e.g. protein-protein interactions).

It is also becoming popular to cross-validate results by running the experiments on a second expression profiling platform. However, this raises the issues of platform concordance and

comparability. Comparisons between platforms have proven problematic for a number of reasons (reviewed in [45]). Each platform may differ in the probes represented, sample preparation methods, hybridization conditions, scanning technology and so forth. For example, our study of 5 different platforms (Affymetrix GeneChip, LongSAGE, LongSAGELite, 'Classic' MPSS and 'Signature' MPSS) demonstrated systematic and random errors resulting in different G+C content sensitivity for each [85]. These biases would influence whether a gene is detected, and if detected, the level of expression measured. These kinds of platform-specific biases will undoubtedly affect the accuracy of measurements and make any platform comparisons imperfect. However, as optimization and standardization methods have improved, platform correlations have transitioned from quite poor to consistently good making cross-platform validation a viable strategy [45].

Conceptually, mRNA levels are used as a surrogate for protein abundance and activity. However, it is not always the case that a change in RNA level translates to a change in protein level (as discussed above). Therefore, validation of changes at the protein level is commonly carried out using techniques such as western blot analysis, ELISA, immunoblot assays, or immunohistochemistry (IHC). The main technique used in this thesis for validation, IHC, will be discussed further. Many of the same issues apply to the others. IHC is a method that uses antibodies (linked to fluorescent molecules or dyes) to stain, identify and thus locate specific protein molecules in tissue sections (usually using a microscope). IHC has been in use for more than 40 years but has become increasingly important as it has moved from a qualitative assay (is protein X present?) to a more quantitative assay (how much of protein X is present?) [86]. There are many issues that affect the accuracy and reproducibility of IHC. For one, the tissue type and duration of fixation can be particularly important for some antigens. Proteins are relatively stable in wax blocks (e.g. paraffin) but can deteriorate quickly once sectioned [87]. Another issue is the method of antigen retrieval, a process used to remove cross-links formed during tissue fixation and improve antibody penetration. This might involve heating the section in an acid buffer or enzymatic treatment. Insufficient antigen retrieval could lead to low-levels of expression being incorrectly scored as negative. Similarly, differences in antibody concentration can strongly affect the results of a study. Seidal *et al.* (2001) actually found that when low antibody concentrations for ERBB2 were used, low expression was associated with poor survival, but the opposite relationship was observed when using a higher concentration [88]. Different antibodies targeting the same antigen can also produce very different results. Different antibodies are raised

in different hosts, using different target antigens (e.g., different splice forms of the protein), can be monoclonal or polyclonal, and represent different isotypes to name just a few of the complicating factors. Finally, once the tissue is actually stained with the antibody of interest, you must still consider the detection method, scoring systems, and cut-off levels. All of the steps outlined above must be optimized for each antibody and tissue type under study. When comparing results between labs it is particularly important to ensure that the same protocols are followed or the comparisons will be meaningless.

A relatively new method of validating multiple microarray findings (i.e., differentially expressed genes of interest) by IHC is through the use of a tissue microarray (TMA). A TMA turns the traditional microarray design on its head. Instead of arraying a large number of genes that are assayed for expression in one sample, many tissue samples are arrayed (e.g., a large cohort of patient tumours) and assayed for expression of a single gene or protein at a time. This allows a gene identified by genome-wide analysis for a relatively small sample set to be quickly validated against a much larger sample population. Figure 1.5 outlines a general protocol for TMA analysis [89]. Briefly, up to 1,000 tissue cores are taken from archived tissue/tumour blocks and arrayed into a recipient 'master block'. This block is then sectioned to produce up to several hundred slides for simultaneous analysis of the patient cohort. The resulting slides can then be analyzed for protein expression, RNA expression or DNA alterations. IHC is commonly used to detect and score all the tissue samples for qualitative or semi-quantitative expression of some protein of interest. This expression can then be related back to clinical and pathological data using standard statistical methods such as survival curves, and categorical association statistics. The combination of conventional expression profiling microarrays and TMAs represent a powerful strategy for the identification and then validation of biologically relevant gene expression changes. The application of this strategy to the problem of human cancer will be discussed further below and in subsequent chapters.

1.5. Gene expression analysis and cancer

In an effort to identify the genes important in cancer, many studies have compared the global expression profiles of cancer and normal tissues [90-97]. Others compare undifferentiated and well differentiated cancers of the same origin to identify genes important in cancer progression [98-108]. Some studies have used cancer expression data to define new subtypes of cancer and develop diagnostic and prognostic indicators. For example, Armstrong *et al.* (2002) have shown

that acute lymphoblastic leukemias (ALL) with a specific translocation have a distinct gene expression profile from other ALL cases that allowed these cases to be classified as a new kind of leukemia, mixed-lineage leukemia (MLL) [109]. Additionally, they identified FLT3 as the most differentially expressed gene in their analysis, a gene previously implicated in leukemogenesis [110, 111]. Such examples demonstrate the potential of cancer expression studies, but in general, genes implicated in cancer by expression data alone are not considered reliable [112]. In most cases, there are dozens or hundreds of genes identified as differentially expressed. Thus, the challenge for many of these studies is to identify which genes are actually directly involved in cancer formation and progression [113]. This section will first discuss the basic mechanisms behind human cancer and then review the use of gene expression analysis to further our understanding of these mechanisms.

1.5.1. Molecular mechanisms of cancer

The transformation of normal cells into cancer cells is one of the most studied and yet still poorly understood processes in biological science. There are more than 200 distinct types of cancer that can be classified into an even greater number of sub-types [114]. Numerous lines of evidence indicate that tumour development is a multi-step process analogous to Darwinian evolution in which successive genetic alterations confer one or another growth advantage. These acquired capabilities can be broken into several broad categories: self-sufficiency in growth signals; insensitivity to antigrowth signals; evasion of apoptosis; limitless replicative ability; sustained angiogenesis; and tissue invasion and metastasis [115] (Figure 1.6). It is the quantity and variability of pathways through these steps that makes studying cancer such a difficult problem. The order in which these capabilities are acquired is extremely variable between and within recognized tumour types. In many cases, a genetic alteration will contribute only partially to an acquired capability. In other cases, the alteration will simultaneously activate several of them. Furthermore, there are many different kinds of alterations that can create these capabilities. However, the most common kind of alteration is probably sequence mutation.

There are two common categories of sequence mutation in tumourigenesis. The first is simple mutation or base substitution. For our purposes, this will include small insertions or deletions, frameshift mutations, missense mutations, and antisense mutations. The second class is chromosomal aberration. This includes large deletions or insertions, translocations, amplifications, and aneuploidy. There is a general belief that the former operates predominantly

in epithelial tumours while the latter operates in haematological and mesenchymal tumours. There are many cases of specific aberrations and gene rearrangements in haematological, bone, and soft tissue tumours whereas these cases are rare in epithelial tumours and the trend has been to focus on gene mutations and deletions for these cancers [116]. However, a recent study suggests that this belief may be based on selective interpretation of the data. Mitelman *et al.* (2004) found that the lower number of recurrent translocations for epithelial tumours was directly related to the number of karyotyped cases available for these cancers [117]. The problem relates to several quantitative and qualitative shortcomings of cytogenetic analysis for these tumours: (1) Chromosome morphology of solid tumours (especially epithelial) is often poor and thus many published solid tumours are only partially karyotyped; (2) Even when good, karyotypes are often so complex that pathogenetically important aberrations can't be distinguished from secondary aberrations; (3) Clonal heterogeneity introduces further complexity. Thus, cytogenetic aberrations resulting in deregulated or fusion genes may be more important in epithelial tumourigenesis than generally believed. The converse is also likely true. Many of the genes disrupted by recurrent translocations may also be vulnerable to smaller mutations in haematological cancers. Until recently (and even now) our ability to comprehensively assess all mutations for a tumour has been severely limited (especially for the small alterations). However, as sequencing becomes more affordable, this situation may improve in the near future.

There are also different modes by which DNA alterations result in cancer. The better characterized mode of action is an alteration of coding sequence that affects the translated protein in some way [112]. This could involve a single amino acid change that alters the protein's function, a truncated or fusion protein, or a complete loss of protein production. In a classic example, a mutation (either inherited or acquired) such as the del(13)(q14) chromosomal deletion or a point mutation leads to the loss of one functional copy of the Retinoblastoma (Rb) gene. A second mutation in the remaining copy of Rb produces a truncated or unstable form of the protein. Rb, a tumour suppressor gene, acts by repressing transcription and its absence leads to uncontrolled cell growth and cancer [118]. Thus, in a single example we see that both base substitutions and chromosomal rearrangements can contribute to the disruption of a gene's coding sequence. In many cases, the observation of these mutations or aberrations has led to the identification of tumour suppressor genes and oncogenes. A recent census identified 291 known cancer genes [112]. The majority were found in leukemias, lymphomas, and sarcomas even

though they represent only 10-15% of human cancer. These genes are usually altered by chromosomal translocations resulting in creation of a fusion gene. For example, the well studied bcr-abl fusion protein results from the Philadelphia Translocation t(9;22) and is considered essential for most cases of chronic myelogenous leukemia (CML) [116]. The fusion results in inappropriate activation of the abl tyrosine kinase and is thought to affect cell-cycle control, apoptosis and cell differentiation.

A second mode of action for cancer-causing alterations involves the regulatory sequences controlling gene expression rather than the gene coding sequence. Both classes of sequence mutations (discussed above) can affect gene regulation and both have been implicated in the deregulation of gene expression in cancer. I will give an illustrative example of each. For chromosomal aberrations, a translocation resulting in deregulation of the MYC oncoprotein is one of the best studied and was the first case identified. A recent study reviewed and analyzed the various mutations observed [119]. Deregulation is caused by a t(8;14)(q24;q32) between the IG locus and MYC that is implicated in most cases of Burkitt's Lymphoma. MYC overexpression results from promoter shift, block of transcription elongation or presence of IG enhancers and is correlated to breakpoint location. Deregulation as a result of base substitutions is less well characterized. This may be due to our limited knowledge of gene regulation rather than the incidence of such events. As mentioned previously, most gene regulatory elements are poorly understood and defined. Thus, a major chromosomal rearrangement that completely substitutes the upstream region of a gene is easier to link to deregulation than small base substitutions. However, there are several convincing examples. A recent study demonstrated a cancer specific mutation in the promoter region of the Survivin (BIRC5) gene [120]. It reports that 68% of cancer-specific cell lines (colon, prostate, and breast cancers) contain a C to G transversion at -31 that was not found in any of the normal cell lines tested. This mutation was found to be within a CDE/CHR transcriptional repressor. A luciferase reporter construct was used to demonstrate that the -31 mutation, as well as other mutation/deletions of the CDE/CHR repressor element, cause increased expression of BIRC5. RT-PCR and western blotting showed a strong correlation between transcript/protein levels and the mutation with high expression in cancer cell lines and low expression in normal lines. BIRC5 encodes a unique inhibitor of apoptosis and plays an important role in both apoptosis and cell cycle regulation and has been reported as abnormally over-expressed in a wide variety of cancers. Thus, the observed mutation

in the Survivin promoter may contribute to over-expression of the anti-apoptosis gene that it encodes and ultimately contribute to development of cancer.

To summarize, mutations are almost universally observed in cancer. In fact, many consider genomic instability as a definitive and requisite characteristic of cancer [115] and, identifying mutations in cancer cells is a common method of defining cancer genes [112]. Furthermore, there are numerous examples where reduced expression of tumour suppressor genes or increased expression of oncogenes is associated with cancer [118]. One of the possible mechanisms for such a change in expression is loss of proper transcriptional regulation. However, few studies have identified mutations that affect a gene's regulatory control rather than its coding sequence. Before this can happen, an improved understanding of human gene regulatory sequences is still needed.

1.5.2. Cancer diagnosis and prognosis using tumour gene expression signatures

As mentioned above, cancer classification systems currently include over 200 types of human cancers and an even greater number of cancer subtypes. Indeed, patient-specific tumour characteristics such as rate of proliferation, capacity for invasion or metastasis, and resistance to specific treatments create an almost continuous spectrum of unique cancers. In order to choose the best treatment, a clinician must determine the cancer type and tumour characteristics as accurately as possible. Molecular methods such as expression profiling arrays represent a powerful new method for improving diagnostic precision. This potential was first demonstrated for diagnosis of acute leukemia. Golub *et al.* (1999) used supervised analysis to identify a diagnostic panel of 50 genes differentially expressed in a test set of 27 acute lymphoblastic leukemia (ALL) cases and 11 acute myeloid leukemia (AML) cases [121]. Using this predictor they were able to diagnose an independent set of 34 leukemias with a high degree of accuracy (29 of 34 correctly classified). Similar methods have since been utilized to predict the development of metastasis and patient prognosis in solid tumours [105, 122]. Lubitz *et al.* (2006) recently used DNA microarray analysis to identify 25 differentially expressed genes in a comparison of 26 benign and 24 malignant thyroid carcinomas [123]. Unsupervised hierarchical clustering was used to classify 22 fine-needle aspirate biopsy (FNAB) specimens. The classification was 100% concordant to the final histological diagnosis compared to 76% from preoperative cytological FNAB diagnosis. For a further review of the use of microarrays for cancer diagnosis and classification see [124].

1.5.3. Defining new molecular subtypes with gene expression data

In addition to discriminating between known subtypes of cancer, expression profiling allows the definition of new cancer subtypes at the molecular level. In combination with clinical data, these new molecular subclasses can provide important information for cancer diagnosis and prognosis. Alizadeh *et al.* (2000) demonstrated this potential for diffuse large B-cell lymphoma (DLBCL) [125]. Using unsupervised class discovery methods, they were able to define two new clinically relevant subgroups of DLBCL termed 'germinal center B-like' and 'activated B-like'. These two new groups have significantly different prognoses five years after chemotherapy treatment with 76% of germinal centre B-like DLBCL patients surviving compared to only 16% for activated B-like DLBCL. Molecular subgroups of clinical relevance have also been established for adult acute myeloid leukemia [126] and breast carcinoma [101, 127]. In thyroid cancer, Giordano *et al.* (2005) performed expression analysis of 51 papillary thyroid carcinomas [128]. They were able to define three tumour groups based on expression patterns that closely reflects tumour morphology and mutational status of BRAF, RAS and RET/PTC.

1.5.4. Cross-platform integration and meta-analyses

To identify genes important in cancer, many studies have compared the global gene expression patterns between normal and cancerous tissue or between different cancer subtypes. Such analyses usually attempt to identify differentially expressed (up- or down-regulated) genes that play an important role in disease development and progression. Unfortunately, these studies tend to identify many genes (dozens or hundreds) of which many are expected to be false-positives and only a small fraction useful as diagnostic or prognostic markers or therapeutic targets. A logical approach for distinguishing important genes from spurious genes, given a large number of candidate gene lists is to search for the intersection of genes identified in multiple independent studies [113]. It is expected that biologically relevant genes will be over-represented and system-specific spurious genes will be under-represented. In these meta-analysis methods, when different platforms or studies are in agreement, there is good evidence for a biological effect. But, when there is disagreement there may not necessarily be good evidence of a non-effect. Thus, meta-analysis of expression profiling minimizes false-positives but may not necessarily minimize false-negatives.

As large numbers of cancer profiling studies have become available the identification of gene list intersections and other meta-analysis methods have become increasingly popular [129]. Rhodes *et al.* (2004) collected and analyzed data from 3,700 cancer samples representing 10 different tumour types [113]. This dataset allowed them to identify a common transcriptional profile universally activated across most cancer types and another signature associated with more aggressive, undifferentiated cancers. Others have focused on specific tumours such as colorectal cancer [130]. One successful example of this technique was used to identify AMACR as a biomarker for distinguishing benign from malignant prostate samples [131]. AMACR was selected using four independent gene expression datasets that all showed it to be over-expressed in prostate cancer. The microarray results were validated at both the RNA and protein level using RT-PCR and immunoblot assay, respectively. Finally, the diagnostic utility of AMACR as a biomarker was confirmed by IHC in TMA studies. It has since gone on to become one of the most useful diagnostic markers for prostate cancer [132]. This illustrates the utility of employing multi-study confirmation and multi-method validation to facilitate biomarker discovery.

Such studies, while conceptually simple, face a number of technical challenges which include: inconsistent gene identifiers; unavailable data; uncertain significance of results; poor quality of probe annotations; and poor description of samples or experimental design. If two or more experiments are carried out in the same lab, combining results can be very powerful. But, if carried out in different labs, much more caution must be exercised. While it can be advantageous to use different technologies, as already discussed, it is less ideal to use different normalization, filtering thresholds, probe mapping techniques, etc. Also, a general advantage of meta-analysis methods is that weak but consistent effects may be detected which would likely be filtered out as uninteresting 'noise' in any individual study. To avoid these limitations, researchers performing meta-analyses need access to raw expression profiling data and high quality experimental and clinical annotations. To this end, all gene expression publications should comply with the MIAME standard and researchers should deposit their raw data in a public database.

1.5.5. Developing biomarkers or panels from microarray class predictors

Another valuable contribution of expression profiling data for cancer diagnoses is for the selection of surrogate molecular markers. Instead of using a microarray gene-signature for diagnosis, promising genes are transferred to a low-throughput technology such as RT-PCR,

ELISA or IHC. This has the potential for a diagnostic test to be developed that is more cost effective and can rapidly be adopted into clinical practice. Also, it may be possible to test for these biomarkers in serum or other bodily fluids thus avoiding invasive diagnostic tests. Biomarkers can be selected from the list of predictors in the microarray classifier. However, these classifiers are often comprised of hundreds of genes which individually may have low predictive power or do not translate well into a new assay. A differentially expressed gene on a microarray will not always be validated by RT-PCR and changes in RNA level do not always correspond to changes at the protein level. Therefore, a major challenge is to identify a gene or panel of genes that retain good predictive power in a low-throughput assay. Using multiple classification algorithms to select genes may help to identify the most robust predictors [133]. Tissue microarrays can be used to test a large number of potential biomarkers using IHC before selecting the final candidates. A comparison of cDNA microarrays and tissue microarrays on 55 breast tumours showed that in many cases (two thirds of the 15 breast cancer markers examined) there was no correlation between the mRNA levels and protein levels [134]. Furthermore, in some cases, the protein levels had prognostic value but not the RNA or vice versa. Therefore, it is a good idea to test a large number of potential markers in a high or medium throughput experiment in order to choose the best candidates for the final low-throughput diagnostic assay for the clinic. A six-gene diagnostic panel (kit, Hs.296031, Hs.24183, LSM7, SYNGR2, and C21or4) was developed in this manner for thyroid cancer. The genes were first demonstrated to have classification potential in a microarray study [135] and then confirmed by qRT-PCR as being able to differentiate between benign and malignant thyroid tumours with high sensitivity and specificity [136].

1.6. Thesis objectives and chapter summaries

Human cancer is a complex disease encompassing numerous pathways and genes. Genome-scale gene expression profiling has become a powerful tool for identifying potential biomarkers, drug targets, and cancer mechanisms. Hundreds or thousands of such studies have been carried out. This has produced an overwhelming amount of data. Processing and making sense of all of these data remains a major bioinformatics challenge. The general aim of this thesis was to develop new computational methods and approaches to help understand these massive gene expression datasets and relate this understanding to gene regulation and cancer. A summary of these approaches as presented in chapters 2 through 5 are given here.

In Chapter 2, I describe an approach for combining multiple gene expression platforms to decrease noise and improve confidence in coexpression predictions. I hypothesized that coexpression relationships confirmed by multiple datasets or technologies would prove more reliable than those identified in a single platform. I compared large publicly available datasets for SAGE, cDNA microarray, and oligonucleotide microarray platforms and found generally poor concordance in their coexpression measures. However, coexpression predictions became more reproducible with larger sample numbers and each of the three platforms performed better as the measure of coexpression (Pearson correlation) increased. Furthermore, gene pairs confirmed by more than one platform (high 2-platform average Pearson) were much more likely to share a Gene Ontology term than those identified by only a single platform. By using the Gene Ontology to choose thresholds I identified a set of high-confidence coexpressed gene pairs for use in regulatory element prediction or other integration studies.

In Chapter 3, I describe the implementation and assessment of a subspace clustering algorithm capable of processing very large gene expression datasets. This continues from work in Chapter 2 where I found that global coexpression measurements were not very reproducible between different platforms. One possible explanation for this is that few genes are actually globally coregulated across all tissues and conditions. Instead, it may be that most genes are coregulated only under more specific conditions. To investigate this possibility, a subspace clustering method was needed to identify groups of genes that are tightly coexpressed under specific subsets of the samples assayed. This presented a major computational challenge because of the nearly infinite number of possible subspaces which must be examined. In fact, none of the existing algorithms were able to process datasets of the size I wanted to investigate. To address this challenge I worked with Byron Gao to develop a new clustering algorithm, called KiWi, which discovers significant subspace clusters from massive datasets. I extensively validated the resulting clusters for these datasets and showed that KiWi correctly assigns redundant probes to the same cluster, groups experiments with common experimental annotations, differentiates real promoter sequences from negative control sequences, and groups genes which share common biological processes and common regulatory sequences.

In Chapter 4, I describe a novel meta-analysis method for combining data from differential expression studies when raw data are unavailable. I applied this method to thyroid cancer and validated some of the predictions on tissue microarrays. This builds on the findings in Chapter 2

where I showed that combining coexpression measurements from multiple platforms increased confidence in coexpression predictions. I hypothesized that the same approach might be useful for differential expression data. However, for most published microarray studies on thyroid cancer, the raw data were not available. Therefore, I developed a simple vote-counting strategy that can be used on published gene lists to rank candidates according to amount of overlap, total sample numbers, and average fold-change. I found a number of statistically significant multi-study genes. These included both well-characterized and entirely novel potential biomarkers for thyroid cancer. Some of the meta-analysis candidates along with a large number of other interesting candidates were tested by IHC on TMAs. These experiments identified a number of promising diagnostic and prognostic markers for thyroid cancer and allowed the development of a diagnostic classification tool for distinguishing benign from malignant lesions with high accuracy.

In Chapter 5, I describe ORegAnno, a web resource for curation of regulatory element sequences from the published literature. In Chapters 2 and 3, I identified high-confidence coexpressed genes for the purpose of identifying coregulated genes, their regulatory element sequences, and ultimately mechanisms for gene deregulation in cancer. However, a major barrier to identifying regulatory elements (either computationally or experimentally) is the lack of positive (and negative) control sequences. To address this challenge I worked with Stephen Montgomery to develop a database and web resource called ORegAnno which allows community-driven curation of the scientific literature. This tool helps to capture the critical details of experiments documenting regulatory element sequences and their binding factors. This has grown into a highly successful resource with hundreds of users, and made important contributions to several other studies.

The methods and tools presented in these chapters are broadly applicable to the study of gene expression, gene regulation, and human cancer. This work specifically contributes to our understanding of gene expression analysis and suggests approaches for combining data from different studies or platforms to improve confidence in coexpression and differential expression predictions. I have also made available two new tools. The first, an algorithm and open-source software package for subspace clustering, is the first of its kind able to handle the truly massive gene expression datasets now commonly available. The second is a database and web interface for community-driven curation of regulatory control sequences.

In addition to the work described in this thesis, I have been involved in several other collaborative projects at the Genome Sciences Centre which are described elsewhere in publications or accepted manuscripts. My high-confidence coexpression predictions (described in Chapter 2) were used by Gordon Robertson, Mikhail Bilenky and others to assist in the discovery of novel regulatory elements as part of a pipeline and database system called cisRED (www.cisred.org) [137]. Similarly, my collection of processed Affymetrix data (also described in Chapter 2) was used by Asim Siddiqui and others in their investigation of sequence biases in large scale gene expression profiling data [85]. The meta-analysis methods developed for thyroid cancer in Chapter 4 were subsequently applied to colon cancer by Simon Chan [138]. I assisted Sam Wiseman, Adrienne Melck and Sher-Ping Leung to apply my tissue microarray data processing and analysis methods (developed for Chapter 4) to other cancers and protein subsets. For example, detailed analysis of the diagnostic and prognostic significance of Type I Growth Factor Receptor family members was performed for colon cancer [139] and thyroid cancer [140]. A similar analysis of cell cycle regulators was performed for thyroid cancer [141]. Through my work on the ORegAnno resource (described in Chapter 5) I assisted Stein Aerts, Casey Bergman, and others in the development of novel text-mining strategies for regulatory sequence annotation [142]; Gordon Robertson and others with the genome-wide profiling of STAT1 binding using the ChIP-seq technique [16]; and Stephen Montgomery with his survey of genomic properties for the detection of regulatory polymorphisms [143]. Finally, I assisted Malachi Griffith and others in the development of a new microarray platform for assessing alternative splicing [144].

Table 1.1. Defining sensitivity, specificity and accuracy from a 2 x 2 table

Consider a classifier that attempts to determine the disease state ("disease" or "no disease") of a group of patients. Each patient has some true disease state and is also assigned a predicted disease state by the classifier. Patients who actually have the disease are either correctly identified ("true positives") or misclassified as "false negatives". Patients who do not have the disease are either correctly identified ("true negatives") or misclassified as "false positives". Sensitivity refers to the percentage of patients who are classified as having the disease among the group of patients who do have the disease. When most patients with the disease are correctly classified, sensitivity approaches 100%. Specificity refers to the percentage of patients who are classified as being disease-free among the group of patients who do not have the disease. When most patients with no disease are correctly classified, specificity approaches 100%. Accuracy refers to the percentage of correct classifications (true positives and true negatives) among all patients. The overall error rate refers to the percentage of misclassifications (false positives and false negatives) among all patients. Ideally, sensitivity, specificity, and accuracy would all be 100%. However, this is rarely the case and therefore, the relative importance of sensitivity and specificity are sometimes balanced against each other depending on the nature of the disease.

Classifier prediction	**True disease state**	
	Disease	**No disease**
Disease	True positive	False positive
No disease	False negative	True negative

Sensitivity = true positives / (true positives + false negatives)
Specificity = true negatives / (true negatives + false positives)
Accuracy = (true positives + true negatives) / total patients
Error rate = (false positives + false negatives) / total patients

Figure 1.1. Comparison of a simple eukaryotic promoter and extensively diversified metazoan regulatory modules

In (a) we see a simple eukaryotic transcriptional unit consisting of a simple core promoter (TATA), upstream activator sequence (UAS) and silencer element spaced within 100–200 bp of the TATA box that is typically found in unicellular eukaryotes. In contrast, (b) illustrates a complex metazoan transcriptional control module (TCM). The TCM is a complex arrangement of multiple clustered enhancer modules interspersed with silencer and insulator elements which can be located 10–50 kb either upstream or downstream of a composite core promoter containing TATA box (TATA), Initiator sequences (INR), and downstream promoter elements (DPE). Figure reprinted by permission from Macmillan Publishers Ltd: Nature. 424(6945): 147-51, copyright 2003 [3].

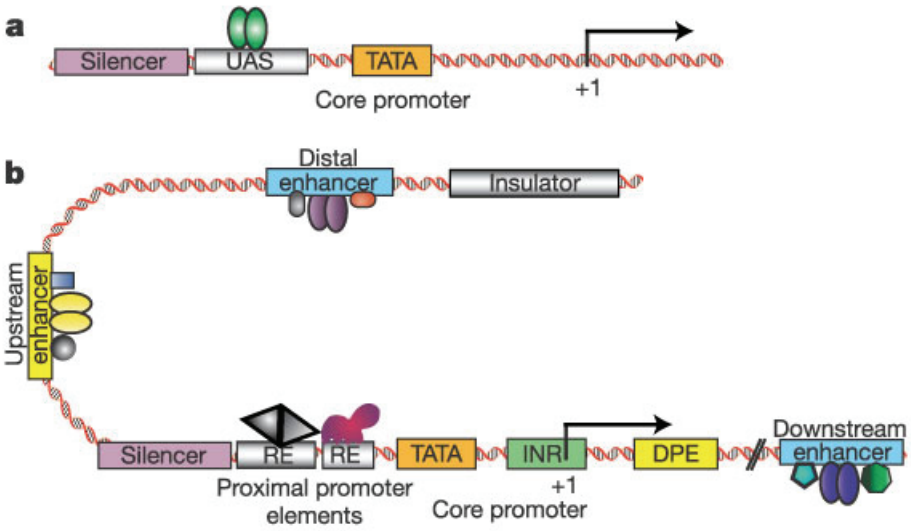


Figure 1.2. The levels of gene expression and gene regulation

The pathway from gene to functional protein in humans is a complex process involving several stages including: chromatin decompaction, initiation, elongation, termination, transcript cleavage, 5' capping, 3' polyadenylation, splicing, mRNA packaging and export (from nucleus to cytoplasm), translation, protein folding, and post-translational modification [5]. While often presented as independent steps, in reality, each stage is part of a continuous process, physically and functionally interconnected. Some stages are regulated by protein (or RNA) factors which in turn are regulated by other factors, extracellular signalling molecules, or environmental cues. Others are regulated by epigenetic modifications such as DNA and histone methylation and acetylation. Much of this regulation involves binding of regulatory factors to DNA regulatory sequences. Figure reprinted from Cell, 108, Orphanides, G. and D. Reinberg, A unified theory of gene expression, pages 439-51, Copyright 2002, with permission from Elsevier [5].

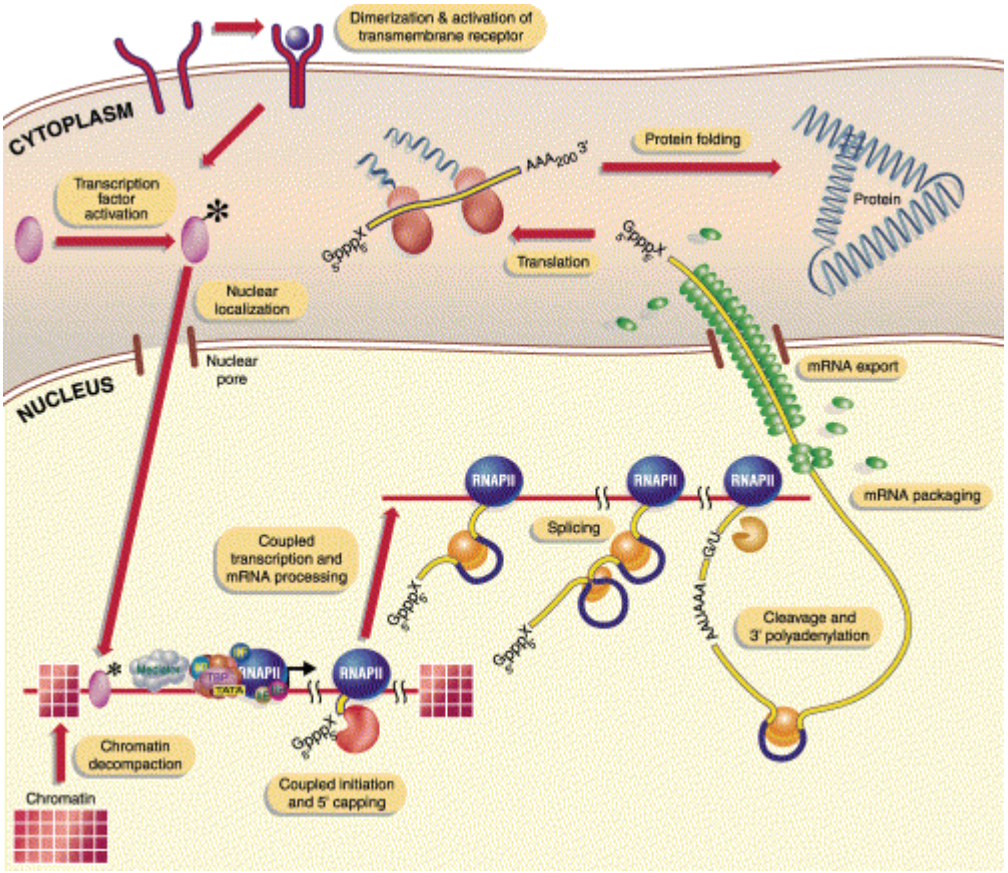
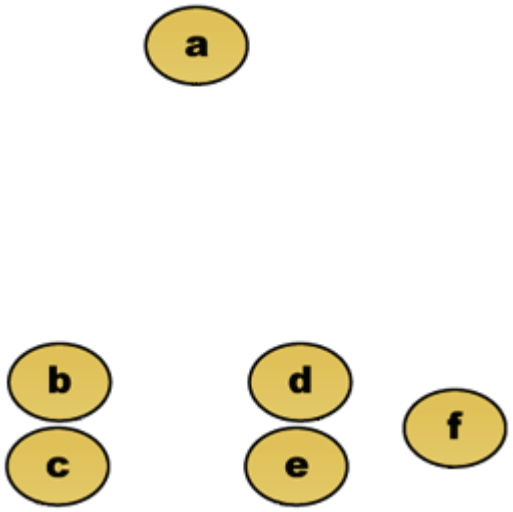


Figure 1.3. A cartoon depiction of how raw data might be clustered using an agglomerative hierarchical clustering method

In (a), raw data are shown for six elements with various geometrical distances between them. These distances could represent a similarity measurement. For example, if the elements represent genes, then the distances might represent a Pearson correlation for their expression patterns across some series of experiments (i.e., highly coexpressed genes are closer to each other). In agglomerative clustering, each element is initially placed in its own group and then these groups are progressively joined based on which are most similar to each other. This process is outlined in (b). In the first step, 'b' and 'c' are grouped together and 'd' and 'e' are grouped together. The elements in both of these pairs are separated by the same distance. Next, the 'de' group is clustered with 'f' as the next most similar group. The process continues until all clusters are grouped into a single super cluster, forming a complete dendrogram. (Figure source: wikipedia, http://en.wikipedia.org/wiki/Cluster_analysis, copyrights released to public domain).

a. raw data



b. agglomerative hierarchical clustering

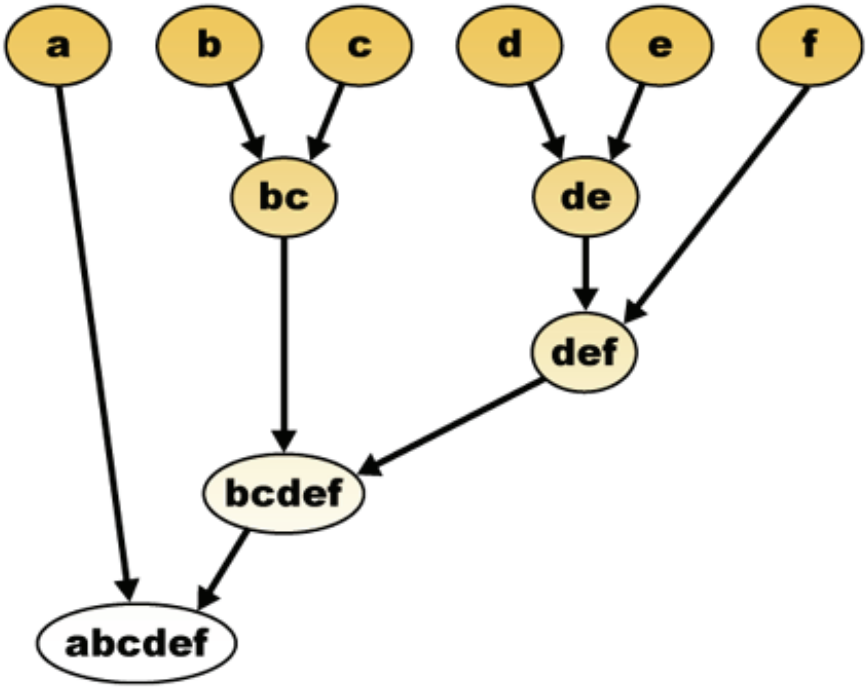


Figure 1.4. Cluster analysis applied to gene expression data

In (A) we see a representation of unordered gene expression data with genes along the rows and experiments along the columns. In (B), the genes are clustered using hierarchical clustering. In (C), the samples are partitioned using k-means clustering (with k=2). Figure reproduced by permission from Quackenbush (2006) [70].

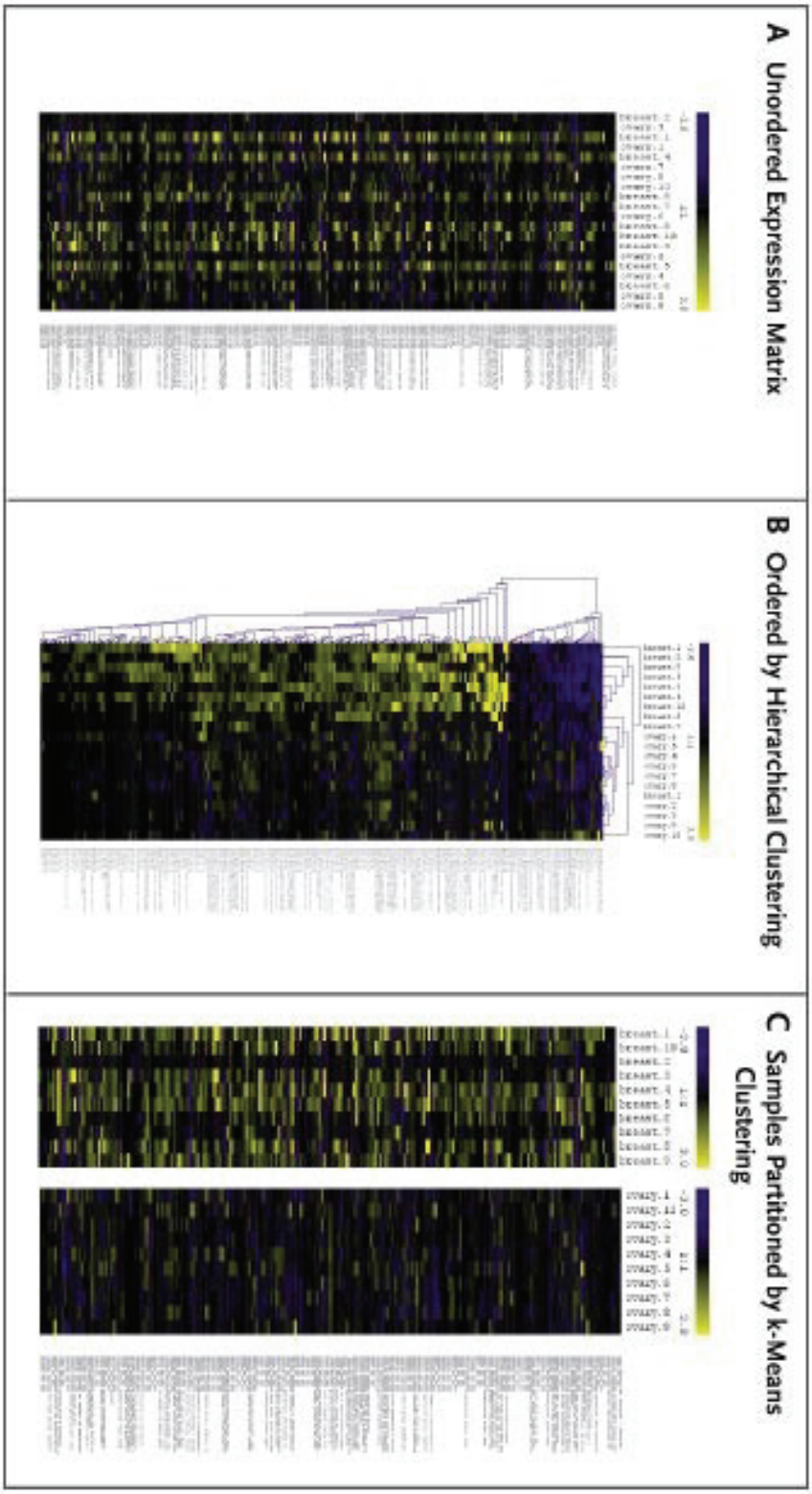


Figure 1.5. A general protocol for tissue microarray (TMA) analysis

Up to 1,000 tissue cores are taken from archived tissue/tumour donor blocks and arrayed into a recipient 'master block'. This block is then sectioned to produce up to several hundred slides for simultaneous analysis of the patient cohort. The resulting slides can be analyzed for protein expression, RNA expression or DNA alterations. Figure reprinted by permission from Macmillan Publishers Ltd: Nature Clinical Practice Oncology. 1(2): 104-11, Copyright 2004 [89].

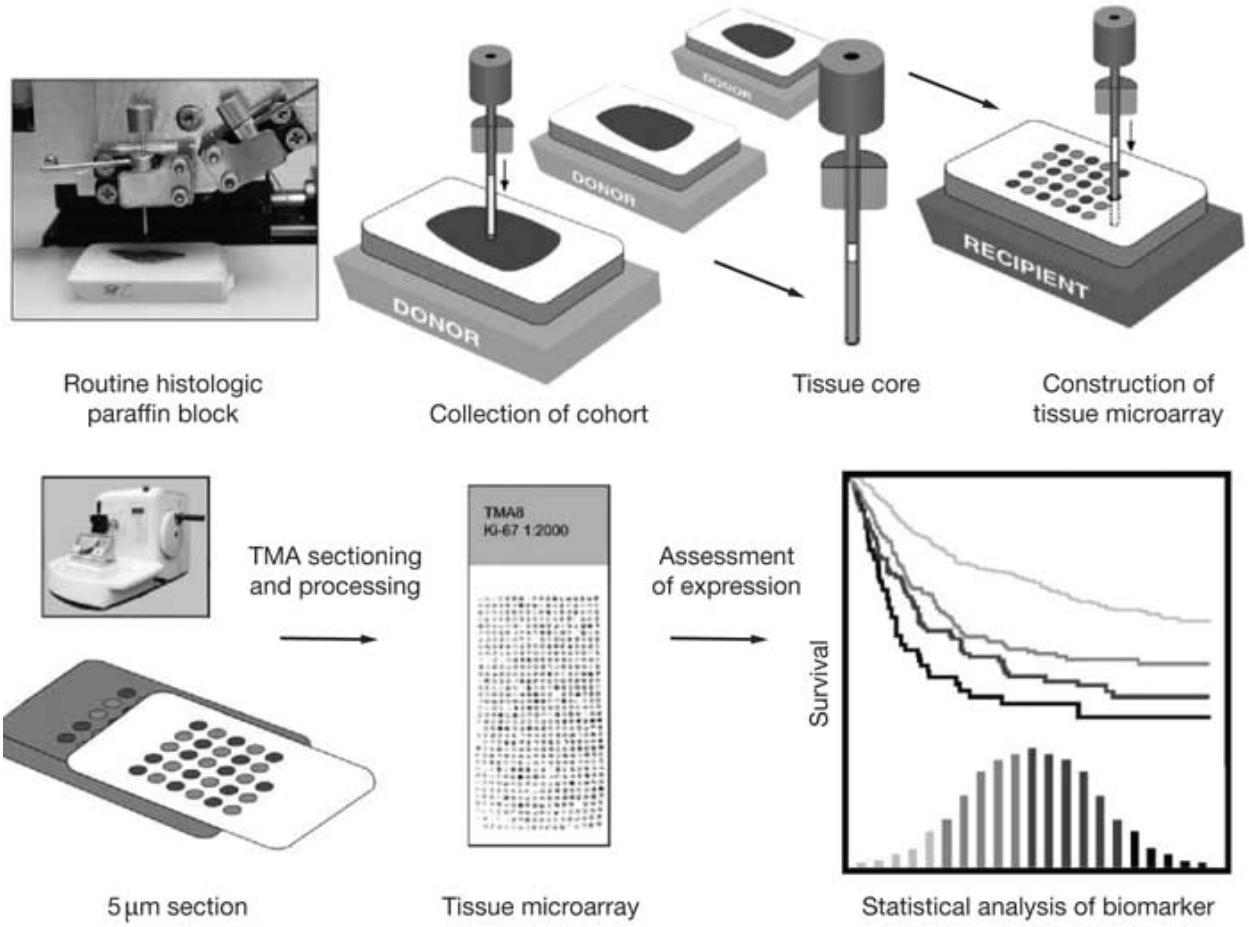
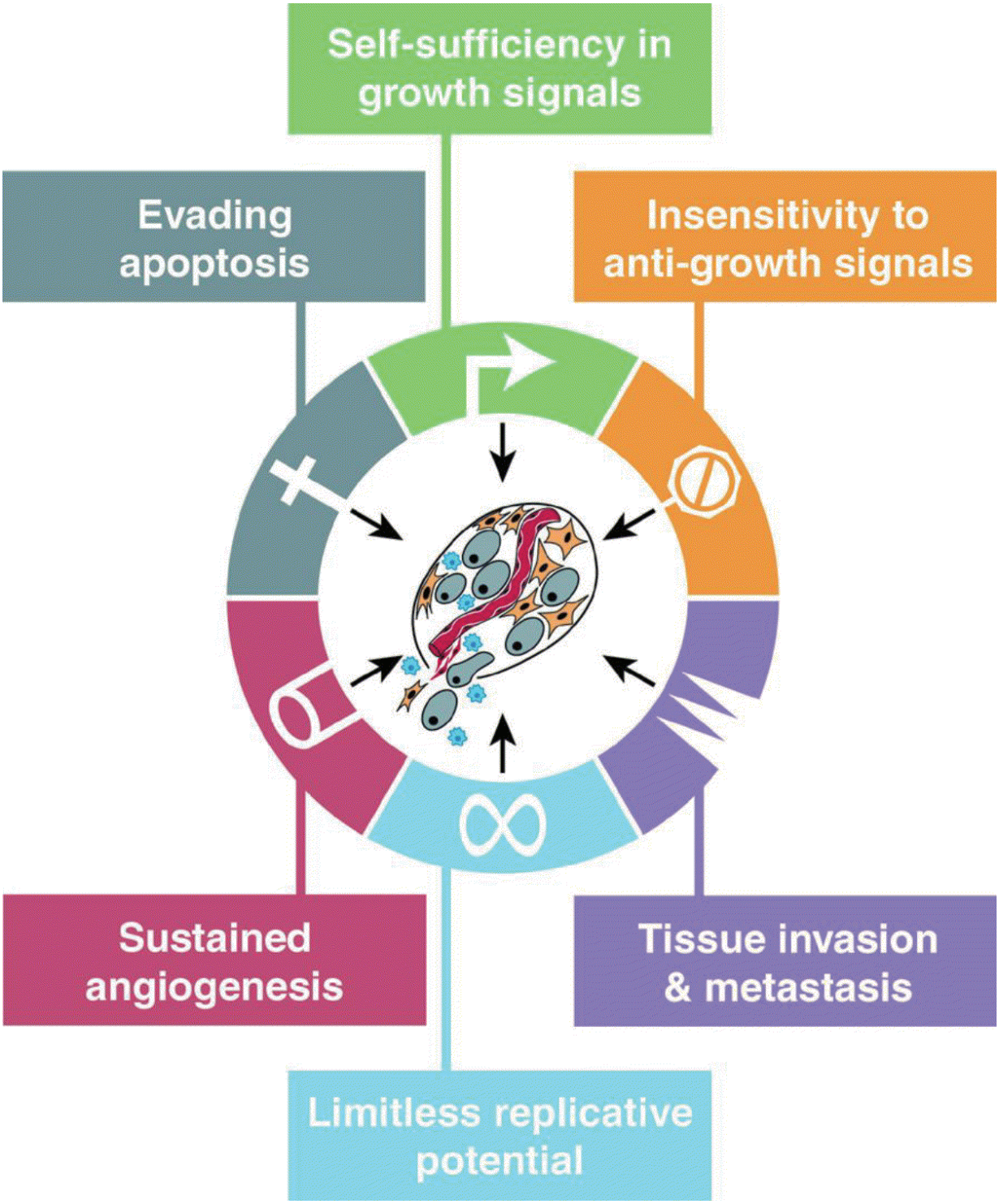


Figure 1.6. Acquired capabilities of cancer

One model for tumour development suggests a multi-step process analogous to Darwinian evolution in which successive genetic alterations confer one or another growth advantage, allowing progressive conversion of normal cells into cancerous cells. These capabilities can be broken into several broad categories: self-sufficiency in growth signals; insensitivity to antigrowth signals; evasion of apoptosis; limitless replicative ability; sustained angiogenesis; and tissue invasion and metastasis. Figure reprinted from Cell, 100, Hanahan, D. and R.A. Weinberg, The hallmarks of cancer, pages 57-70, Copyright 2000, with permission from Elsevier [115].



References

1. International Human Genome Sequencing Consortium, I.H.G.S.C., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
2. Pennisi, E., *Searching for the genome's second code.* Science, 2004. **306**(5696): p. 632-5.
3. Levine, M. and R. Tjian, *Transcription regulation and animal diversity.* Nature, 2003. **424**(6945): p. 147-51.
4. Berman, B.P., B.D. Pfeiffer, T.R. Laverty, S.L. Salzberg, G.M. Rubin, M.B. Eisen, and S.E. Celniker, *Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura.* Genome Biol, 2004. **5**(9): p. R61.
5. Orphanides, G. and D. Reinberg, *A unified theory of gene expression.* Cell, 2002. **108**(4): p. 439-51.
6. Narlikar, G.J., H.Y. Fan, and R.E. Kingston, *Cooperation between complexes that regulate chromatin structure and transcription.* Cell, 2002. **108**(4): p. 475-87.
7. Richards, E.J. and S.C. Elgin, *Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects.* Cell, 2002. **108**(4): p. 489-500.
8. Arnone, M.I. and E.H. Davidson, *The hardwiring of development: organization and function of genomic regulatory systems.* Development, 1997. **124**(10): p. 1851-64.
9. Hack, C.J., *Integrated transcriptome and proteome data: the challenges ahead.* Brief Funct Genomic Proteomic, 2004. **3**(3): p. 212-9.
10. Elnitski, L., V.X. Jin, P.J. Farnham, and S.J. Jones, *Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques.* Genome Res, 2006. **16**(12): p. 1455-64.
11. Wasserman, W.W. and A. Sandelin, *Applied bioinformatics for the identification of regulatory elements.* Nat Rev Genet, 2004. **5**(4): p. 276-87.
12. Tompa, M., N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, W.J. Kent, V.J. Makeev, A.A. Mironov, W.S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu, *Assessing computational tools for the discovery of transcription factor binding sites.* Nat Biotechnol, 2005. **23**(1): p. 137-44.
13. Vlieghe, D., A. Sandelin, P.J. De Bleser, K. Vleminckx, W.W. Wasserman, F. van Roy, and B. Lenhard, *A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.* Nucleic Acids Res, 2006. **34**(Database issue): p. D95-7.
14. Matys, V., O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, and E. Wingender, *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* Nucleic Acids Res, 2006. **34**(Database issue): p. D108-10.
15. Ren, B., F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T.L. Volkert, C.J. Wilson, S.P. Bell, and R.A. Young, *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.
16. Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O.L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones, *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.

17. Wei, C.L., Q. Wu, V.B. Vega, K.P. Chiu, P. Ng, T. Zhang, A. Shahab, H.C. Yong, Y. Fu, Z. Weng, J. Liu, X.D. Zhao, J.L. Chew, Y.L. Lee, V.A. Kuznetsov, W.K. Sung, L.D. Miller, B. Lim, E.T. Liu, Q. Yu, H.H. Ng, and Y. Ruan, *A global map of p53 transcription-factor binding sites in the human genome.* Cell, 2006. **124**(1): p. 207-19.
18. Ng, P., J.J. Tan, H.S. Ooi, Y.L. Lee, K.P. Chiu, M.J. Fullwood, K.G. Srinivasan, C. Perbost, L. Du, W.K. Sung, C.L. Wei, and Y. Ruan, *Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes.* Nucleic Acids Res, 2006. **34**(12): p. e84.
19. Schena, M., D. Shalon, R.W. Davis, and P.O. Brown, *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
20. Lockhart, D.J., H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown, *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.
21. Velculescu, V.E., L. Zhang, B. Vogelstein, and K.W. Kinzler, *Serial analysis of gene expression.* Science, 1995. **270**(5235): p. 484-7.
22. Brenner, S., M. Johnson, J. Bridgham, G. Golda, D.H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S.R. Williams, K. Moon, T. Burcham, M. Pallas, R.B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran, *Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.* Nat Biotechnol, 2000. **18**(6): p. 630-4.
23. Ahmed, F.E., *Molecular techniques for studying gene expression in carcinogenesis.* J Environ Sci Health C Environ Carcinog Ecotoxicol Rev, 2002. **20**(2): p. 77-116.
24. Cowell, J.K. and L. Hawthorn, *The application of microarray technology to the analysis of the cancer genome.* Curr Mol Med, 2007. **7**(1): p. 103-20.
25. Hermeking, H., *Serial analysis of gene expression and cancer.* Curr Opin Oncol, 2003. **15**(1): p. 44-9.
26. Luo, J., W.B. Isaacs, J.M. Trent, and D.J. Duggan, *Looking beyond morphology: cancer gene expression profiling using DNA microarrays.* Cancer Invest, 2003. **21**(6): p. 937-49.
27. Mandruzzato, S., *Technological platforms for microarray gene expression profiling.* Adv Exp Med Biol, 2007. **593**: p. 12-8.
28. Mazumder, A. and Y. Wang, *Gene-expression signatures in oncology diagnostics.* Pharmacogenomics, 2006. **7**(8): p. 1167-73.
29. Petersen, D.W. and E.S. Kawasaki, *Manufacturing of microarrays.* Adv Exp Med Biol, 2007. **593**: p. 1-11.
30. Pusztai, L., *Chips to bedside: incorporation of microarray data into clinical practice.* Clin Cancer Res, 2006. **12**(24): p. 7209-14.
31. Ramaswamy, S. and T.R. Golub, *DNA microarrays in clinical oncology.* J Clin Oncol, 2002. **20**(7): p. 1932-41.
32. Venkatasubbarao, S., *Microarrays--status and prospects.* Trends Biotechnol, 2004. **22**(12): p. 630-7.
33. Pleasance, E.D. and S.J.M. Jones, *Evaluation of SAGE tags for transcriptome study*, in *SAGE Technologies: Current Technologies and Applications*, S.M. Wang, Editor. 2005, Horizon Bioscience: Norwich, UK.
34. Bentley, D.R., *Whole-genome re-sequencing.* Curr Opin Genet Dev, 2006. **16**(6): p. 545-52.

35. Shendure, J., G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, and G.M. Church, *Accurate multiplex polony sequencing of an evolved bacterial genome.* Science, 2005. **309**(5741): p. 1728-32.
36. Bainbridge, M.N., R.L. Warren, M. Hirst, T. Romanuik, T. Zeng, A. Go, A. Delaney, M. Griffith, M. Hickenbotham, V. Magrini, E.R. Mardis, M.D. Sadar, A.S. Siddiqui, M.A. Marra, and S.J. Jones, *Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach.* BMC Genomics, 2006. **7**: p. 246.
37. Quackenbush, J., *Using DNA microarrays to assay gene expression*, in *Bioinformatics: A practical guide to the analysis of genes and proteins.*, A.D. Baxevanis and B.F.F. Ouellette, Editors. 2005, Wiley-Interscience: Hoboken, NJ. p. 409-444.
38. Cope, L.M., R.A. Irizarry, H.A. Jaffee, Z. Wu, and T.P. Speed, *A benchmark for Affymetrix GeneChip expression measures.* Bioinformatics, 2004. **20**(3): p. 323-31.
39. Lee, M.L., F.C. Kuo, G.A. Whitmore, and J. Sklar, *Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations.* Proc Natl Acad Sci U S A, 2000. **97**(18): p. 9834-9.
40. Mocellin, S. and C.R. Rossi, *Principles of gene microarray data analysis.* Adv Exp Med Biol, 2007. **593**: p. 19-30.
41. Wu, Z., R.A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer, *A Model-Based Background Adjustment for Oligonucleotide Expression Arrays.* Journal of the American Statistical Association, 2004. **99**: p. 909-17.
42. Ahmed, F.E., *Microarray RNA transcriptional profiling: part II. Analytical considerations and annotation.* Expert Rev Mol Diagn, 2006. **6**(5): p. 703-15.
43. Porter, D.A., I.E. Krop, S. Nasser, D. Sgroi, C.M. Kaelin, J.R. Marks, G. Riggins, and K. Polyak, *A SAGE (serial analysis of gene expression) view of breast tumor progression.* Cancer Res, 2001. **61**(15): p. 5697-702.
44. Verhaak, R.G., F.J. Staal, P.J. Valk, B. Lowenberg, M.J. Reinders, and D. de Ridder, *The effect of oligonucleotide microarray data pre-processing on the analysis of patient-cohort studies.* BMC Bioinformatics, 2006. **7**: p. 105.
45. Yauk, C.L. and M.L. Berndt, *Review of the literature examining the correlation among DNA microarray technologies.* Environ Mol Mutagen, 2007. **48**(5): p. 380-94.
46. Harbig, J., R. Sprinkle, and S.A. Enkemann, *A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array.* Nucleic Acids Res, 2005. **33**(3): p. e31.
47. Diehn, M., G. Sherlock, G. Binkley, H. Jin, J.C. Matese, T. Hernandez-Boussard, C.A. Rees, J.M. Cherry, D. Botstein, P.O. Brown, and A.A. Alizadeh, *SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data*. Nucleic Acids Res, 2003. **31**(1): p. 219-23.
48. Dennis, G., Jr., B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, and R.A. Lempicki, *DAVID: Database for Annotation, Visualization, and Integrated Discovery.* Genome Biol, 2003. **4**(5): p. P3.
49. Bouton, C.M. and J. Pevsner, *DRAGON: Database Referencing of Array Genes Online.* Bioinformatics, 2000. **16**(11): p. 1038-9.
50. Tsai, J., R. Sultana, Y. Lee, G. Pertea, S. Karamycheva, V. Antonescu, J. Cho, B. Parvizi, F. Cheung, and J. Quackenbush, *RESOURCERER: a database for annotating and linking microarray resources within and across species.* Genome Biol, 2001. **2**(11): p. SOFTWARE0002.
51. Lash, A.E., C.M. Tolstoshev, L. Wagner, G.D. Schuler, R.L. Strausberg, G.J. Riggins, and S.F. Altschul, *SAGEmap: a public gene expression resource.* Genome Res, 2000. **10**(7): p. 1051-60.

52. Robertson, N., M. Oveisi-Fordorei, S.D. Zuyderduyn, R.J. Varhol, C. Fjell, M. Marra, S. Jones, and A. Siddiqui, *DiscoverySpace: an interactive data analysis application.* Genome Biol, 2007. **8**(1): p. R6.
53. Mecham, B.H., G.T. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D.Z. Wetmore, T.J. Mariani, I.S. Kohane, and Z. Szallasi, *Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements.* Nucleic Acids Res, 2004. **32**(9): p. e74.
54. Audic, S. and J.M. Claverie, *The significance of digital gene expression profiles.* Genome Res, 1997. **7**(10): p. 986-95.
55. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
56. Steinhoff, C. and M. Vingron, *Normalization and quantification of differential expression in gene expression microarrays.* Brief Bioinform, 2006. **7**(2): p. 166-77.
57. Ruijter, J.M., A.H. Van Kampen, and F. Baas, *Statistical evaluation of SAGE libraries: consequences for experimental design.* Physiol Genomics, 2002. **11**(2): p. 37-44.
58. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J R Stat Soc [Ser B], 1995. **57**(1): p. 289-300.
59. Pounds, S.B., *Estimation and control of multiple testing error rates for microarray studies.* Brief Bioinform, 2006. **7**(1): p. 25-36.
60. Swift, S., A. Tucker, V. Vinciotti, N. Martin, C. Orengo, X. Liu, and P. Kellam, *Consensus clustering and functional interpretation of gene-expression data.* Genome Biol, 2004. **5**(11): p. R94.
61. Belacel, N., Q. Wang, and M. Cuperlovic-Culf, *Clustering methods for microarray gene expression data.* Omics, 2006. **10**(4): p. 507-31.
62. Kim, S.K., J. Lund, M. Kiraly, K. Duke, M. Jiang, J.M. Stuart, A. Eizinger, B.N. Wylie, and G.S. Davidson, *A gene expression map for Caenorhabditis elegans.* Science, 2001. **293**(5537): p. 2087-92.
63. Jelinsky, S.A., P. Estep, G.M. Church, and L.D. Samson, *Regulatory networks revealed by transcriptional profiling of damaged Saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes.* Mol Cell Biol, 2000. **20**(21): p. 8157-67.
64. Allocco, D.J., I.S. Kohane, and A.J. Butte, *Quantifying the relationship between co-expression, co-regulation and gene function.* BMC Bioinformatics, 2004. **5**(1): p. 18.
65. Quackenbush, J., *Genomics. Microarrays--guilt by association.* Science, 2003. **302**(5643): p. 240-1.
66. Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet, 2003. **34**(2): p. 166-76.
67. Ihmels, J., S. Bergmann, and N. Barkai, *Defining transcription modules using large-scale gene expression data.* Bioinformatics, 2004. **20**(13): p. 1993-2003.
68. Simonis, N., S.J. Wodak, G.N. Cohen, and J. van Helden, *Combining pattern discovery and discriminant analysis to predict gene co-regulation.* Bioinformatics, 2004. **20**(15): p. 2370-9.
69. Fan, J. and Y. Ren, *Statistical analysis of DNA microarray data in cancer research.* Clin Cancer Res, 2006. **12**(15): p. 4469-73.
70. Quackenbush, J., *Microarray analysis and tumor classification.* N Engl J Med, 2006. **354**(23): p. 2463-72.
71. de Hoon, M.J., S. Imoto, J. Nolan, and S. Miyano, *Open source clustering software.* Bioinformatics, 2004. **20**(9): p. 1453-4.

72. Frank, E., M. Hall, L. Trigg, G. Holmes, and I.H. Witten, *Data mining in bioinformatics using Weka.* Bioinformatics, 2004. **20**(15): p. 2479-81.
73. Gewehr, J.E., M. Szugat, and R. Zimmer, *BioWeka--extending the Weka framework for bioinformatics.* Bioinformatics, 2007. **23**(5): p. 651-3.
74. Saeed, A.I., N.K. Bhagabati, J.C. Braisted, W. Liang, V. Sharov, E.A. Howe, J. Li, M. Thiagarajan, J.A. White, and J. Quackenbush, *TM4 microarray software suite.* Methods Enzymol, 2006. **411**: p. 134-93.
75. Schadt, E.E., C. Li, B. Ellis, and W.H. Wong, *Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.* J Cell Biochem Suppl, 2001. **Suppl 37**: p. 120-5.
76. Barrett, T., T.O. Suzek, D.B. Troup, S.E. Wilhite, W.C. Ngau, P. Ledoux, D. Rudnev, A.E. Lash, W. Fujibuchi, and R. Edgar, *NCBI GEO: mining millions of expression profiles--database and tools.* Nucleic Acids Res, 2005. **33**(Database issue): p. D562-6.
77. Sherlock, G., T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J.C. Matese, S.S. Dwight, M. Kaloper, S. Weng, H. Jin, C.A. Ball, M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, and J.M. Cherry, *The Stanford Microarray Database.* Nucleic Acids Res, 2001. **29**(1): p. 152-5.
78. Brazma, A., H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G.G. Lara, A. Oezcimen, P. Rocca-Serra, and S.A. Sansone, *ArrayExpress--a public repository for microarray gene expression data at the EBI.* Nucleic Acids Res, 2003. **31**(1): p. 68-71.
79. Rhodes, D.R., J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A.M. Chinnaiyan, *ONCOMINE: a cancer microarray database and integrated data-mining platform.* Neoplasia, 2004. **6**(1): p. 1-6.
80. Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* Nat Genet, 2001. **29**(4): p. 365-71.
81. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
82. Kelso, J., J. Visagie, G. Theiler, A. Christoffels, S. Bardien, D. Smedley, D. Otgaar, G. Greyling, C.V. Jongeneel, M.I. McCarthy, T. Hide, and W. Hide, *eVOC: a controlled vocabulary for unifying gene expression data.* Genome Res, 2003. **13**(6A): p. 1222-30.
83. Kudla, K.M. and M.C. Rallins, *SNOMED: a controlled vocabulary for computer-based patient records.* J Ahima, 1998. **69**(5): p. 40-4; quiz 45-6.
84. Provenzano, M. and S. Mocellin, *Complementary techniques: validation of gene expression data by quantitative real time PCR.* Adv Exp Med Biol, 2007. **593**: p. 66-73.
85. Siddiqui, A.S., A.D. Delaney, A. Schnerch, O.L. Griffith, S.J. Jones, and M.A. Marra, *Sequence biases in large scale gene expression profiling data.* Nucleic Acids Res, 2006. **34**(12): p. e83.
86. Walker, R.A., *Quantification of immunohistochemistry--issues concerning methods, utility and semiquantitative assessment I.* Histopathology, 2006. **49**(4): p. 406-10.

87. Bertheau, P., D. Cazals-Hatem, V. Meignin, A. de Roquancourt, O. Verola, A. Lesourd, C. Sene, C. Brocheriou, and A. Janin, *Variability of immunohistochemical reactivity on stored paraffin slides.* J Clin Pathol, 1998. **51**(5): p. 370-4.
88. Seidal, T., A.J. Balaton, and H. Battifora, *Interpretation and quantification of immunostains.* Am J Surg Pathol, 2001. **25**(9): p. 1204-7.
89. Giltnane, J.M. and D.L. Rimm, *Technology insight: Identification of biomarkers with tissue microarray technology.* Nat Clin Pract Oncol, 2004. **1**(2): p. 104-11.
90. Alon, U., N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, *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**(12): p. 6745-50.
91. Luo, J., D.J. Duggan, Y. Chen, J. Sauvageot, C.M. Ewing, M.L. Bittner, J.M. Trent, and W.B. Isaacs, *Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling.* Cancer Res, 2001. **61**(12): p. 4683-8.
92. Luo, J.H., Y.P. Yu, K. Cieply, F. Lin, P. Deflavia, R. Dhir, S. Finkelstein, G. Michalopoulos, and M. Becich, *Gene expression analysis of prostate cancers.* Mol Carcinog, 2002. **33**(1): p. 25-35.
93. Magee, J.A., T. Araki, S. Patil, T. Ehrig, L. True, P.A. Humphrey, W.J. Catalona, M.A. Watson, and J. Milbrandt, *Expression profiling reveals hepsin overexpression in prostate cancer.* Cancer Res, 2001. **61**(15): p. 5692-6.
94. Notterman, D.A., U. Alon, A.J. Sierk, and A.J. Levine, *Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays.* Cancer Res, 2001. **61**(7): p. 3124-30.
95. Welsh, J.B., P.P. Zarrinkar, L.M. Sapinoso, S.G. Kern, C.A. Behling, B.J. Monk, D.J. Lockhart, R.A. Burger, and G.M. Hampton, *Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer.* Proc Natl Acad Sci U S A, 2001. **98**(3): p. 1176-81.
96. Frierson, H.F., Jr., A.K. El-Naggar, J.B. Welsh, L.M. Sapinoso, A.I. Su, J. Cheng, T. Saku, C.A. Moskaluk, and G.M. Hampton, *Large scale molecular analysis identifies genes with altered expression in salivary adenoid cystic carcinoma.* Am J Pathol, 2002. **161**(4): p. 1315-23.
97. Garber, M.E., O.G. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, M. van de Rijn, G.D. Rosen, C.M. Perou, R.I. Whyte, R.B. Altman, P.O. Brown, D. Botstein, and I. Petersen, *Diversity of gene expression in adenocarcinoma of the lung.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13784-9.
98. Singh, D., P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, and W.R. Sellers, *Gene expression correlates of clinical prostate cancer behavior.* Cancer Cell, 2002. **1**(2): p. 203-9.
99. Ye, Q.H., L.X. Qin, M. Forgues, P. He, J.W. Kim, A.C. Peng, R. Simon, Y. Li, A.I. Robles, Y. Chen, Z.C. Ma, Z.Q. Wu, S.L. Ye, Y.K. Liu, Z.Y. Tang, and X.W. Wang, *Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning.* Nat Med, 2003. **9**(4): p. 416-23.
100. van 't Veer, L.J., H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend, *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-6.
101. Sorlie, T., C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, T. Thorsen, H. Quist, J.C. Matese, P.O. Brown, D.

Botstein, P. Eystein Lonning, and A.L. Borresen-Dale, *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.
102. Shipp, M.A., K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, and T.R. Golub, *Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.* Nat Med, 2002. **8**(1): p. 68-74.
103. Schwartz, D.R., S.L. Kardia, K.A. Shedden, R. Kuick, G. Michailidis, J.M. Taylor, D.E. Misek, R. Wu, Y. Zhai, D.M. Darrah, H. Reed, L.H. Ellenson, T.J. Giordano, E.R. Fearon, S.M. Hanash, and K.R. Cho, *Gene expression in ovarian cancer reflects both morphology and biological behavior, distinguishing clear cell from other poor-prognosis ovarian carcinomas.* Cancer Res, 2002. **62**(16): p. 4722-9.
104. Rickman, D.S., M.P. Bobek, D.E. Misek, R. Kuick, M. Blaivas, D.M. Kurnit, J. Taylor, and S.M. Hanash, *Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis.* Cancer Res, 2001. **61**(18): p. 6885-91.
105. Ramaswamy, S., K.N. Ross, E.S. Lander, and T.R. Golub, *A molecular signature of metastasis in primary solid tumors.* Nat Genet, 2003. **33**(1): p. 49-54.
106. Dyrskjot, L., T. Thykjaer, M. Kruhoffer, J.L. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, and T.F. Orntoft, *Identifying distinct classes of bladder carcinoma using microarrays.* Nat Genet, 2003. **33**(1): p. 90-6.
107. Chen, X., S.T. Cheung, S. So, S.T. Fan, C. Barry, J. Higgins, K.M. Lai, J. Ji, S. Dudoit, I.O. Ng, M. Van De Rijn, D. Botstein, and P.O. Brown, *Gene expression patterns in human liver cancers.* Mol Biol Cell, 2002. **13**(6): p. 1929-39.
108. Bhattacharjee, A., W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J. Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub, D.J. Sugarbaker, and M. Meyerson, *Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13790-5.
109. Armstrong, S.A., J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, and S.J. Korsmeyer, *MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.* Nat Genet, 2002. **30**(1): p. 41-7.
110. Zhao, M., H. Kiyoi, Y. Yamamoto, M. Ito, M. Towatari, S. Omura, T. Kitamura, R. Ueda, H. Saito, and T. Naoe, *In vivo treatment of mutant FLT3-transformed murine leukemia with a tyrosine kinase inhibitor.* Leukemia, 2000. **14**(3): p. 374-8.
111. Tse, K.F., G. Mukherjee, and D. Small, *Constitutive activation of FLT3 stimulates multiple intracellular signal transducers and results in transformation.* Leukemia, 2000. **14**(10): p. 1766-76.
112. Futreal, P.A., L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M.R. Stratton, *A census of human cancer genes.* Nat Rev Cancer, 2004. **4**(3): p. 177-83.
113. Rhodes, D.R., J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A.M. Chinnaiyan, *Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.* Proc Natl Acad Sci U S A, 2004. **101**(25): p. 9309-14.
114. Greene, F.L., D.L. Page, I.D. Fleming, A. Fritz, C.M. Balch, D.G. Haller, and M. Morrow, eds. *AJCC Cancer Staging Manual*. 6th ed. 2002, Springer-Verlag: New York, NY.
115. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.

116. Rowley, J.D., *Chromosome translocations: dangerous liaisons revisited.* Nat Rev Cancer, 2001. **1**(3): p. 245-50.
117. Mitelman, F., B. Johansson, and F. Mertens, *Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer.* Nat Genet, 2004. **36**(4): p. 331-4.
118. Macdonald, F., C.H.J. Ford, and A.G. Casson, *Molecular Biology of Cancer*. 2 ed. Advanced Text. 2004, New York: BIOS Scientific Publishers. 269.
119. Wilda, M., K. Busch, I. Klose, T. Keller, W. Woessmann, J. Kreuder, J. Harbott, and A. Borkhardt, *Level of MYC overexpression in pediatric Burkitt's lymphoma is strongly dependent on genomic breakpoint location within the MYC locus.* Genes Chromosomes Cancer, 2004. **41**(2): p. 178-82.
120. Xu, Y., F. Fang, G. Ludewig, G. Jones, and D. Jones, *A mutation found in the promoter region of the human survivin gene is correlated to overexpression of survivin in cancer cells.* DNA Cell Biol, 2004. **23**(9): p. 527-37.
121. Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.
122. Roepman, P., L.F. Wessels, N. Kettelarij, P. Kemmeren, A.J. Miles, P. Lijnzaad, M.G. Tilanus, R. Koole, G.J. Hordijk, P.C. van der Vliet, M.J. Reinders, P.J. Slootweg, and F.C. Holstege, *An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas.* Nat Genet, 2005. **37**(2): p. 182-6.
123. Lubitz, C.C., S.K. Ugras, J.J. Kazam, B. Zhu, T. Scognamiglio, Y.T. Chen, and T.J. Fahey, 3rd, *Microarray analysis of thyroid nodule fine-needle aspirates accurately classifies benign and malignant lesions.* J Mol Diagn, 2006. **8**(4): p. 490-8; quiz 528.
124. Perez-Diez, A., A. Morgun, and N. Shulzhenko, *Microarrays for cancer diagnosis and classification.* Adv Exp Med Biol, 2007. **593**: p. 74-85.
125. Alizadeh, A.A., M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, and L.M. Staudt, *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**(6769): p. 503-11.
126. Bullinger, L., K. Dohner, E. Bair, S. Frohling, R.F. Schlenk, R. Tibshirani, H. Dohner, and J.R. Pollack, *Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia.* N Engl J Med, 2004. **350**(16): p. 1605-16.
127. Perou, C.M., T. Sorlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S.X. Zhu, P.E. Lonning, A.L. Borresen-Dale, P.O. Brown, and D. Botstein, *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
128. Giordano, T.J., R. Kuick, D.G. Thomas, D.E. Misek, M. Vinco, D. Sanders, Z. Zhu, R. Ciampi, M. Roh, K. Shedden, P. Gauger, G. Doherty, N.W. Thompson, S. Hanash, R.J. Koenig, and Y.E. Nikiforov, *Molecular classification of papillary thyroid carcinoma: distinct BRAF, RAS, and RET/PTC mutation-specific gene expression profiles discovered by DNA microarray analysis.* Oncogene, 2005. **24**(44): p. 6646-56.
129. Cahan, P., A.M. Ahmad, H. Burke, S. Fu, Y. Lai, L. Florea, N. Dharker, T. Kobrinski, P. Kale, and T.A. McCaffrey, *List of lists-annotated (LOLA): A database for annotation and comparison of published microarray gene lists.* Gene, 2005. **360**(1): p. 78-82.
130. Shih, W., R. Chetty, and M.S. Tsao, *Expression profiling by microarrays in colorectal cancer (Review).* Oncol Rep, 2005. **13**(3): p. 517-24.

131. Rubin, M.A., M. Zhou, S.M. Dhanasekaran, S. Varambally, T.R. Barrette, M.G. Sanda, K.J. Pienta, D. Ghosh, and A.M. Chinnaiyan, *alpha-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer.* Jama, 2002. **287**(13): p. 1662-70.
132. Jiang, Z., B.A. Woda, C.L. Wu, and X.J. Yang, *Discovery and clinical application of a novel prostate cancer marker: alpha-methylacyl CoA racemase (P504S).* Am J Clin Pathol, 2004. **122**(2): p. 275-89.
133. Fu, L.M. and C.S. Fu-Liu, *Multi-class cancer subtype classification based on gene expression signatures with reliability analysis.* FEBS Lett, 2004. **561**(1-3): p. 186-90.
134. Ginestier, C., E. Charafe-Jauffret, F. Bertucci, F. Eisinger, J. Geneix, D. Bechlian, N. Conte, J. Adelaide, Y. Toiron, C. Nguyen, P. Viens, M.J. Mozziconacci, R. Houlgatte, D. Birnbaum, and J. Jacquemier, *Distinct and complementary information provided by use of tissue and DNA microarrays in the study of breast tumor markers.* Am J Pathol, 2002. **161**(4): p. 1223-33.
135. Mazzanti, C., M.A. Zeiger, N.G. Costouros, C. Umbricht, W.H. Westra, D. Smith, H. Somervell, G. Bevilacqua, H.R. Alexander, and S.K. Libutti, *Using gene expression profiling to differentiate benign versus malignant thyroid tumors.* Cancer Res, 2004. **64**(8): p. 2898-903.
136. Rosen, J., M. He, C. Umbricht, H.R. Alexander, A.P. Dackiw, M.A. Zeiger, and S.K. Libutti, *A six-gene model for differentiating benign from malignant thyroid tumors on the basis of gene expression.* Surgery, 2005. **138**(6): p. 1050-6; discussion 1056-7.
137. Robertson, G., M. Bilenky, K. Lin, A. He, W. Yuen, M. Dagpinar, R. Varhol, K. Teague, O.L. Griffith, X. Zhang, Y. Pan, M. Hassel, M.C. Sleumer, W. Pan, E.D. Pleasance, M. Chuang, H. Hao, Y.Y. Li, N. Robertson, C. Fjell, B. Li, S.B. Montgomery, T. Astakhova, J. Zhou, J. Sander, A.S. Siddiqui, and S.J. Jones, *cisRED: a database system for genome-scale computational discovery of regulatory elements.* Nucleic Acids Res, 2006. **34**(Database issue): p. D68-73.
138. Chan, S.K., O.L. Griffith, I.T. Tai, and S.J.M. Jones, *Meta-analysis of Colorectal Cancer Gene Expression Profiling Studies Identifies Consistently Reported Candidate Biomarkers.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(3): p. 543–52.
139. Leung, S., O.L. Griffith, H. Masoudi, A. Gown, S.J.M. Jones, T. Phang, and S.M. Wiseman, *Clinical Utility of Type I Growth Factor Receptor Expression by Colon Cancer.* Am J Surg, 2008. **In press.**
140. Wiseman, S.M., O.L. Griffith, A. Melck, H. Masoudi, A. Gown, R. Nabi, and S.J.M. Jones, *Evaluation of Type 1 Growth Factor Receptor Family Expression in 205 Thyroid Lesions Reveals Diagnostic Utility and Targeted Therapeutic Potential for HER1, HER3, and HER4.* Am J Surg, 2008. **In press.**
141. Melck, A., H. Masoudi, O.L. Griffith, A. Rajput, G. Wilkins, S. Bugis, S.J. Jones, and S.M. Wiseman, *Cell Cycle Regulators Show Diagnostic and Prognostic Utility for Differentiated Thyroid Cancer.* Ann Surg Oncol, 2007. **14**(12): p. 3403-11.
142. Aerts, S., M. Haeussler, O.L. Griffith, S. Van Vooren, S.J.M. Jones, S.B. Montgomery, C.M. Bergman, and T.O.R.A. Consortium, *Text-mining assisted regulatory annotation.* Genome Biol, 2008. **9**(2): p. R31.1-13.
143. Montgomery, S.B., O.L. Griffith, J.M. Schuetz, A. Brooks-Wilson, and S.J. Jones, *A survey of genomic properties for the detection of regulatory polymorphisms.* PLoS Comput Biol, 2007. **3**(6): p. e106.
144. Griffith, M., M.J. Tang, O.L. Griffith, S.Y. Chan, J.K. Asano, T. Zeng, S. Flibotte, A. Ally, A. Baross, R.D. Morin, M. Hirst, S.J.M. Jones, G.B. Morin, I.T. Tai, and M.A. Marra, *ALEXA – A microarray design platform for alternative expression analysis.* Nat Methods, 2008. **5**(2): p. 118.

2. Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses [3,4]

2.1. Introduction

Large-scale expression profiling has become an important tool for the identification of gene functions, disease mechanisms, and regulatory elements. The development of three such techniques, cDNA microarrays [1], oligonucleotide (oligo) microarrays [2] and serial analysis of gene expression (SAGE) [3] has resulted in a plethora of studies attempting to elucidate cellular processes by identifying groups of genes that appear to be coexpressed. Our motivation for this study was to explore the fecundity of large extant expression datasets to identify coexpressed genes and their utility as a resource for biological study. Coexpression data are increasingly used for validation and integration with other 'omic' data sources such as sequence conservation [4], yeast two-hybrid interactions [5, 6], RNA interference [7] and regulatory element predictions [8] to name only a few. If different platforms or datasets produce widely different measures of coexpression it could have significant impacts on the results of such studies. Furthermore, methods to assess these datasets and identify a coherent, consistent picture of coexpression will be needed.

High degrees of consistency within a platform ($r > 0.9$ for both technical and biological replicates) have been reported for cDNA microarrays, Affymetrix oligonucleotide microarrays, and SAGE [9-12]. However, it should be noted that good correlations for SAGE data are dependent to a great deal on sufficient sampling depth, especially for low-abundance tags [13]. Cross-platform comparisons of gene expression values have found 'reasonable' correlations for matched samples, especially for more highly expressed transcripts [11, 14-20]. Other comparisons have reported 'poor' correlations [16, 19, 21-25]. Yauk and Berndt (2007) recently reviewed the multitude of cross-platform comparison studies and concluded that improvements in the technologies, analysis methods and annotations have generally led to much higher levels of correlations in recent years (2004 to 2007) [26]. The correlations reported above were for expression levels or expression changes of individual genes, not coexpression of gene pairs. To

[3] A portion of this chapter has been published. Griffith O.L., Pleasance E.D., Fulton D.L., Oveisi M., Ester M., Siddiqui A.S., Jones S.J.M. 2005. Assessment and Integration of Publicly Available SAGE, cDNA Microarray, and Oligonucleotide Microarray Expression Data for Global Coexpression Analyses. Genomics. 86:476-488.

[4] Co-authorship details: I was the primary author of Griffith *et al*. (2005) and was responsible for all analysis, text, figures and tables included in this chapter except where indicated below. Erin Pleasance helped to design, perform and write text for the internal consistency analysis (sections 2.2.5 and 2.3.1). Debra Fulton helped to design, perform and write text for the gene ontology analysis (sections 2.2.9 and 2.3.5). Mehrdad Oveisi and Asim Siddiqui assisted with the processing and analysis of SAGE data (section 2.2.2) and contributed invaluable discussion. Martin Ester and Steven Jones supervised and funded the project.

our knowledge, only one study has examined the correlation of coexpression results from multiple platforms. Lee *et al.* (2003) compared matched Affymetrix oligonucleotide chips and spotted cDNA microarrays for the NCI-60 cancer cell lines panel [27]. For each platform, the calculation involved determining the Pearson correlation (r) between expression profiles (across 60 cell lines) for all pairwise gene combinations. Then, a correlation of correlations (r_c) between the two platforms was determined. When all gene pairs were considered a global concordance of $r_c = 0.25$ was reported. As the correlation cutoff was increased, r_c improved steadily to 0.92 at a correlation cutoff of r = 0.91 (but only 28 of 2,061 genes remained). Thus, for most gene pairs there was poor correlation of correlations for global coexpression values.

Genome wide coexpression analyses in *C. elegans*, *S. cerevisiae*, and numerous bacterial species have been used with some success to identify gene functions or genes that are coregulated [28-31]. This "guilt-by-association" approach has received criticism because of high levels of noise and other problems inherent to the methods but still holds great interest for biologists [32, 33]. If matched samples display questionable levels of consistency between expression profiles generated by different platforms the question remains as to how effectively unmatched samples from many different sources will compare. If two genes are coregulated (i.e. controlled by an identical set of transcription factors) they should display similar expression patterns across many conditions and be identified as coexpressed. This is the basic premise of many gene function and regulation studies. If true, large datasets from different expression platforms should identify the same coexpressed gene pairs even if derived from different conditions and tissues. However, it may be that few genes are globally coregulated and thus datasets comprised of different samples will identify different sets of coregulated genes. Similarly, noise and biases inherent to the different methods may result in highly discordant measures of coexpression, even for genes with similar function or under similar regulatory control.

The purpose of this study was to assess the differences between publicly available expression data for global coexpression analyses and investigate the value of combining multiple platforms to decrease noise and improve confidence in coexpression predictions. We have compared large publicly available datasets for SAGE, cDNA microarray (cDNA), and Affymetrix oligonucleotide microarray (Affymetrix) platforms (Figure 2.1). We calculated all gene-to-gene Pearson correlation coefficients and assessed the platforms for internal consistency, cross-platform concordance, and agreement with the Gene Ontology. The Pearson correlation was

chosen as a similarity metric because it is one of the most commonly used, with numerous published examples for Affymetrix [9, 34, 35], cDNA [5, 29, 36] and SAGE [37, 38]. Because the datasets represent unmatched samples, a direct comparison of platforms was challenging. Our results indicate that the three platforms identify very different measures of coexpression for most gene pairs with a very low correlation of correlations between platforms. However, coexpression predictions become more reproducible with larger sample numbers and each of the three platforms performs better (identifies more gene pairs with common GO terms) as the gene pair Pearson correlation increases. Furthermore, gene pairs confirmed by more than one platform (high 2-platform average Pearson) were much more likely to share a GO term than those identified by only a single platform. Other recently published coexpression methods (TMM, ArrayProspector) also performed well against GO at higher scores but identified very different gene pairs. By using the Gene Ontology to choose thresholds of high-confidence pairs for each we identify a set of coexpressed gene pairs that represents the best of each approach.

2.2. Methods

2.2.1. Data sources

Human gene expression data for three major expression platforms were collected from public sources. We used a recently published data set of 1,202 cDNA microarray experiments [4] representing 13,595 genes; 242 SAGE libraries from the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) representing 15,426 genes; and 667 Affymetrix HG-U133A oligonucleotide microarray experiments (889 were available but 667 had PMA detection calls) representing 8,106 genes, also from GEO (Figure 2.1). cDNA microarray genes were identified by LocusLink ids. Therefore this identifier was used for the other two platforms to allow the gene intersection of the three datasets to be determined and used for the subsequent analyses.

2.2.2. Data filtering

cDNA microarray data for 13,595 genes were used as provided by Stuart *et al*. (2003) except for minor formatting changes. The 242 SAGE libraries ranged from 1,430 to 308,589 total tags in size with an average size of 52,723. SAGE data were first filtered to remove tags with less than one count in at least 10 libraries reducing the unique tags from 609,224 to 87,521 (and total tags from 12,758,981 to 11,219,373). Next, SAGE tags were mapped to genes by the 'lowest' sense-strand tag predicted from Refseq [39] or MGC [40] sequences and then mapped to LocusLink ids using the DiscoverySpace software package [41] reducing the tag set further to 47,263 unique

tags. Generally, the 'lowest tag' corresponded to the canonical, 3'-most NlaIII anchoring enzyme recognition site (position 1) expected for the gene sequence. However, if such a canonical match was not found, higher position (less 3') mappings were also accepted. Of the 5,881 genes we mapped tags to and used in the analysis, 2,374 (40%) had tags from the 1st position, 878 (15%) from the 2nd, 586 (10%) from the 3rd, 400 (7%) from the 4th, 335 (6%) from the 5th, and so on. In total, 40% were canonical tags, and 65% were from the lowest 3 positions. It is estimated that 30-50% of human genes are likely to use multiple different polyadenylation signals [42, 43] and 38% of human genes generate more than one tag due to alternative splicing [44]. If these different transcript forms are not all represented in RefSeq or MGC, some of these would appear to be internal tags. Therefore, it is expected that some tags will not map to the canonical 3'-most position and we felt the best approach was to use the mapping that was closest. In the event of discrepancy between Refseq and MGC, the former was taken as correct because a larger number of tags could be mapped with this resource (9,568 vs 6,295) and was thus perceived to be more complete. Only 297 tag types (~5%) with disagreements between Refseq and MGC are represented in the final gene set used in the analysis. Of these, we anticipate that some will be correct mappings, others incorrect, and others ambiguous (i.e. the tag does not uniquely represent a single gene). Thus, in less than 5% of the genes, there may be a bias towards lower correlation for comparisons involving the SAGE data. We believe this had a minimal effect on our overall results and does not alter our conclusions. If a tag mapped to more than one LocusLink or more than one tag mapped to the same LocusLink it was discarded resulting in a final set of 15,426 unique tags (2,762,500 total tags) confidently mapped to LocusLink ids. 22,215 Affymetrix probe ids were mapped to 20,577 LocusLink Ids using the most current Affymetrix annotation file for the HG-U133A chip (www.affymetrix.com, Suppl. Materials). As with the SAGE tags, probes with ambiguous mapping to LocusLink were discarded resulting in a final set of 8,106 genes from the Affymetrix dataset. Once LocusLink ids were available for all three platforms, the intersection was determined. This subset of 5,881 genes, present in all three platforms, was used for all subsequent analyses. The final 5,881 unique SAGE tags represent 1,173,430 total tags sequenced.

2.2.3. Distance calculations

Ratio values for the cDNA microarray data were used as provided for the Pearson calculation. Affymetrix probe intensities were converted to natural log (ln) values. All ln(intensity) values were normalized by subtracting the median and dividing by the inter-quartile range for the

experiment as previously described [45]. Only Affymetrix probe intensities with a present (P) call were considered (p-value < 0.04). Intensities with absent (A) or marginal (M) calls were set to null. To compensate for different library sizes SAGE tag counts were normalized to 10,000-tags/library and log-transformed as follows [38]:

Tag frequency = ln((tag count x 10,000)/total tags in library).

It can be argued that an observation of zero tag counts represents a true measurement of non-expression of the transcript that the tag represents. However, given the large range in library sizes (see above), zero tag counts may also represent insufficient sampling of a low expression-level transcript. We performed distance calculations for the SAGE dataset with both null and zero representing zero tag counts. In the latter case, the vast majority of values are zero. This frequently resulted in potentially spurious high correlations when two genes shared the same value (zero) across nearly all libraries and then one or two non-zero values. The Pearson correlation distribution was extremely skewed towards high positive correlations (data not shown). When nulls were used instead of zero, such tag pairs were excluded based on insufficient minimum common experiments (see below) and the overall correlation distribution was approximately normal. This was much more comparable to the other platforms. Therefore, SAGE tag counts of zero were converted to nulls for all subsequent analyses.

In all platforms, genes are represented by a vector of expression values for all the experiments in the data set. In each case, genes have null values if not represented on that array (cDNA), no tags observed (SAGE), or intensity not significantly detected (Affymetrix). Thus, when calculating Pearson correlations between gene pairs, the number of shared data points varied from zero to the total number of experiments. A minimum number of common experiments (MCE) were required for each gene pair to provide some confidence in the value calculated (a Pearson correlation based on observations from only two experiments is meaningless). A range of MCEs was used for the internal consistency analysis (see below) and then one minimum chosen for subsequent analyses.

A Pearson correlation coefficient was calculated for all possible gene pairs for each platform as a measure of expression similarity. These calculations were performed by a modified version of the C clustering library [46] on 64-bit opteron linux machines with 8-32GB memory. Please see

supplementary web page for modified C source code and explanation of changes (http://www.bcgsc.ca/gc/bomge/coexpression/suppl_materials).

2.2.4. Correlation of correlations analysis

Correlation of correlations (r_c) for internal consistencies and platform comparisons were performed as previously described [27] using the Pearson correlation function (cor) of the R statistical package (version 1.8.1). This correlation involves millions of data points and thus can not be graphed easily. Therefore, data were binned and density plots created using the Bioconductor hexbin library (version 1.0.3) for the R statistical programming language [47].

2.2.5. Internal consistency analysis

To evaluate the consistency of coexpression observed within each platform, we divided the experiments available and determined coexpression for each subset independently. If a platform consistently finds coexpressed genes regardless of the exact experiments involved, the r_c will be close to 1. To determine whether the observed r_c is significant, we repeat the procedure with randomly permuted gene expression values, expecting a r_c close to 0.

2.2.5.1. Pseudo-random division method

Division was performed first randomly, and then pseudo-randomly. The pseudo-random division was necessary to prevent artificially high internal consistencies resulting from comparing mostly replicates (or very similar experiments) in the two subsets. In many cases (especially for the Affymetrix data) experimental replicates or very similar samples exist in the dataset. The purpose of coexpression analysis is to identify genes that behave similarly across many conditions. The internal consistency analysis is meant to measure how consistently a series of experiments across different conditions would identify the same coexpressed genes. If the two subsets of experiments contain replicates, they are more likely to identify the same coexpressed genes as the expression values of the replicates will be very similar. The cross-platform comparisons do not have this advantage because they consist of different experiments. Thus, to make the internal consistency calculation more comparable to the cross-platform comparisons, we used a pseudo-random division for subsequent analysis. Experiments were randomly divided into two subsets but experiments belonging to the same experimental series (Affymetrix), publication (cDNA), or tissue (SAGE) were required to fall into the same subset.

2.2.5.2. Minimum common experiments analysis

Differences in the number of common experiments between any two genes result from missing values in the data matrices. In the case of the cDNA microarray data, different arrays were used in different experiments, and not all genes are present on all the arrays. For SAGE, a tag is often observed in one library but will have a zero tag count in other libraries. For Affymetrix oligonucleotide arrays, an intensity is always reported for every probe but in some cases the Affymetrix statistical software will determine that the probe was not reliably detected and assign an absent (A) or marginal (M) call instead of a present (P) call for that probe. As missing SAGE tags and probes not called Present represent genes expressed below the detection threshold of the SAGE and Affymetrix array experiments, we did not include these data in our analysis. Thus, for each dataset, there were gene pairs that were rarely represented in the same experiment and their Pearson correlations were based on very few data points. The effect of number of common experiments on internal consistency was determined by calculating the internal consistency for a series of datasets with different minimum common experiment (MCE) criteria. 100 different pseudo-random divisions were performed to get an average internal consistency for each MCE threshold. An MCE was chosen for each such that the same internal consistency would result ($r = 0.25$) (Figure 2.2). Thus, all subsequent analyses were based on a MCE cutoff of 95 for Affymetrix, 28 for cDNA, and 23 for SAGE. Requiring a MCE cutoff removes gene pairs from the datasets. To maintain an unbiased comparison, only the 1,173,330 gene pairs common to all three platform datasets (after application of MCE criteria) were used in the subsequent platform comparisons.

2.2.6. Cancer sample analysis

The proportion of cancer samples was determined from the literature for the cDNA dataset [4] and from GEO sample records for Affymetrix and SAGE. SAGE, having the highest percentage of cancer samples, was used for the analysis. The SAGE data set was manually divided into 94 normal and 148 cancer libraries based on sample descriptions from the GEO sample records. The consistency between these two subsets of the data was calculated as described above and compared to the internal consistency results.

2.2.7. Cross-platform correlation analysis

As with the internal consistency analysis, a correlation of gene correlations was calculated, but was determined for each of the three pairwise platform comparisons instead of between subsets

of one platform. If the two platforms being compared report the same correlation between each gene pair, we expect the overall correlation between platforms would be near 1. The global concordance (r_c) was also determined for increasing gene correlation cutoffs to compare to results obtained in the Lee *et al.* (2003) NCI-60 study [27].

2.2.8. Ranked match analysis

In addition to considering the actual Pearson correlation between each gene pair and comparing between platforms, the correlation rank was considered. This analysis identifies shared co-expressed genes, or matches, between platforms. For instance, a shared match would be illustrated by the following: Gene A's 2nd most similar gene is gene B in the Affymetrix data. This is gene A's 3rd most similar gene in the SAGE data. This example would count as one shared 'match' for a neighbourhood of k = 3 for the Affymetrix versus SAGE comparison. A Perl script was written to determine each gene's closest k neighbours from one dataset and compare to another dataset. Numbers of shared neighbours within each neighbourhood size (k) were tallied and graphed. 1,000 randomizations were conducted for each platform comparison to determine how often the level of agreement at each neighbourhood would be observed by chance.

2.2.9. Gene ontology analysis

The Gene Ontology (GO) is a controlled vocabulary that describes the roles of genes and proteins in all organisms [48]. GO is composed of three independent ontologies: biological process, molecular function, and cellular component. The GO descriptive terms are represented as nodes connected by directed edges that may have more than one parent node (directed acyclic graph). A gene is annotated to its most specific GO term description and all ancestor GO terms are implied.

The Gene Ontology (GO) MySQL database dump (release 200402 of assocdb) was downloaded from http://www.godatabase.org/dev/database. A GO MySQL database was built and a Perl script was developed to extract three GO information subspaces from the biological process ontology: 1) the most specific GO terms for each gene; 2) the most specific terms along with their associated parent terms; and 3) the most specific terms along with their associated parent and grandparent terms. Two categories of annotations were used for the evaluation of each GO information subspace: 1) gene annotations that did not include those derived from inferred

electronic annotations (IEAs) (1,007 genes found in common with our data set) and 2) gene annotations including IEAs (1,426 genes found in common with our data set). IEA is used for annotations that depend directly on computation or automated transfer of annotations from a database (i.e., where no curator has checked the annotation to verify its accuracy). Similar results were obtained for both non-IEA and IEA analyses. Only the IEA results are reviewed in the figures and text. A total of 301,536 gene pairs were available with both coexpression data (in all three platforms) and GO annotations (including IEAs) for the subsequent GO analyses.

One potential issue with our analysis is that of a circular argument. It is possible that a coexpressed gene pair could be found to share a common GO term that was annotated for both genes by a coexpression analysis. Thus, coexpression data could be confirming coexpression data. To check for this problem we assessed the degree to which our dataset depends on annotations inferred from expression profiles (IEP evidence code). Only 93 of 32,669 biological process annotations use IEP evidence, corresponding to only 73 genes with one or more IEP annotations. Of these, only 1 was present in our gene set and this gene also had non-IEP annotations. Therefore the potential for a circular argument is negligible.

Results shown in Figure 2.3 were extracted from the gene pair correlation data, by enumerating the number of gene pairs found at common GO terms across a gene's expression similarity neighbourhood for each GO information subspace. To illustrate for the "most specific" subspace and a neighbourhood size of 10 (k=10), for each gene, its 10 most highly coexpressed genes were determined according to their Pearson correlations. For these 10 gene pairs, the single most specific GO term for which they were annotated was retrieved for each gene. Finally, the number of gene pairs with the same most specific GO term was tallied. This process was repeated for all genes in the overlapping gene set. Results shown in Figure 2.4 were extracted by enumerating the number of gene pairs found at common GO terms for each range of Pearson correlations from 0 to 1 in increments of 0.1. The results summarized in Figure 2.5 were enumerated in a similar manner but used average Pearson correlations between two platforms instead of individual Pearson correlations. 1,000 random permutations of the data were conducted to determine how often GO confirmation of a gene pair at each neighbourhood or Pearson range would occur by chance. Scripts were written in Perl and are available at: http://www.bcgsc.ca/gc/bomge/coexpression/suppl_materials.

2.2.10. Comparison to other coexpression methods

Results shown in Figure 2.6 were generated using the GO analysis method described above for Figure 2.4 and Figure 2.5. ArrayProspector (AP) data were obtained by request from the author [49]. Only pairs with scores above 0.150 were provided. TMM data was downloaded from the authors' supplemental webpage [50]. Both negative and positive correlations were included and thus a gene pair can appear twice. Only pairs with absolute scores of 1 or greater were provided. The 2-platform combination (2PC) method represents all 2-platform averages (Affymetrix/cDNA, Affymetrix/SAGE, and cDNA/SAGE). Thus, a gene pair can appear as many as three times if all three pairwise averages fall within the 0-1 range graphed. All datasets were converted from their respective identifiers to Uniprot [51] and the percent of gene pairs found at common GO terms for each range of scores determined. The top 2,500 gene pairs were examined to determine the overlap in results for high scoring pairs. Thresholds for a high-confidence set of coexpressed gene pairs were chosen for each method at the approximate respective score where performance was at least 3 to 4 times better than random chance (2PC > 0.65; AP > 0.7; TMM > 7).

2.2.11. Supplementary materials

Supplementary data files can be found at:
http://www.bcgsc.ca/gc/bomge/coexpression/suppl_materials

2.3. Results

2.3.1. Internal consistency

Before performing cross-platform comparisons, we decided to evaluate each platform individually to determine if different experiments from the same technology identified consistent patterns of gene coexpression. To this end, internal consistency was determined by dividing each of the datasets in half and comparing the gene-to-gene Pearson correlations for each subset (Figure 2.2A-C). We first divided the data in a purely random fashion. However, to make the internal consistency calculation more comparable to the cross-platform comparisons, we also devised a pseudo-random division which takes into account the presence of experimental replicates and very similar experimental conditions in the datasets (see methods).

Internal consistency was found to be dependent on the minimum number of common experiments (MCE) between any two genes on which Pearson correlations were calculated. MCE was defined as follows:

MCE – The minimum number of common or shared experiments for which any two genes actually have values available in their respective expression profiles (Figure 2.2D).

Increasing the MCE cutoff increased the internal consistency but decreased the number of gene pairs considered for both the pseudo-random (Figure 2.2) and random (Figure 2.7) division methods. With the random division, and an MCE of 100, Affymetrix showed the highest average internal correlation of 0.925, then cDNA microarray with correlation of 0.889, and SAGE with correlation of 0.776. This MCE cutoff was also used by the group that provided the cDNA microarray data [4] (E. Segal, pers. comm.). As expected, the pseudo-random division, which groups replicates and experimental datasets, reduced internal consistencies with values of 0.253 for Affymetrix, 0.273 for the cDNA microarray and 0.660 for SAGE with MCE of 100 (Figure 2.2). Unfortunately, as the SAGE dataset contains only 242 samples, division into two groups of approximately 120 results in relatively few gene pairs that meet the criteria of 100 MCE (only 1,518 pairs on average). Although approximately 60% of these SAGE libraries are derived from cancer samples, we found no evidence of an effect on the coexpression results (Figure 2.8) and therefore included them in subsequent analysis (see section 2.3.2 for more details)

Internal consistency is a measure of the reproducibility or robustness of gene coexpression predictions, similar to a cross-validation test. This is based on the assumption that if a gene pair is truly coexpressed based on an expression dataset, it should be predicted as coexpressed by random subsets of the data. The consistency increases with higher MCE but at different rates for the three datasets because of their different natures in terms of number of experiments and experiment composition. Thus, it would be unfair to compare the datasets with MCEs that resulted in different levels of reproducibility. Studies generally choose some cutoff for a minimum number of common experiments such as 5, 10 or 100 [4, 34, 50]. In an effort to produce an unbiased comparison of the three platforms, the pseudorandom division was used to determine an appropriate MCE which would generate the same internal consistency ($r_c = 0.25$)

for each (Affymetrix MCE = 95; cDNA MCE = 28; SAGE MCE = 23) (Figure 2.2). All internal consistency correlations are summarized in Table 2.1.

2.3.2. Cancer sample analysis

Cancer samples were found to represent a substantial fraction in the cDNA (~29%), Affymetrix (~40% of the complete 889 samples) and SAGE (~61%) datasets. Cancer tissues are often characterized by changes in gene expression and thus could act as a confounding factor when trying to identify coexpressed genes. To investigate this issue the SAGE dataset was divided into cancer and normal subsets and consistency between these measured. The comparison of normal and cancer SAGE libraries resulted in a correlation of 0.324 for an MCE of 23 and 0.707 for an MCE of 80 (MCE of 100 could not be used because the normal tissue subset only contained 94 samples). These results are comparable to that seen for consistencies of SAGE when not taking cancer status into account (Figure 2.8). Thus, we cautiously concluded that the presence of cancer libraries was not seriously affecting the SAGE co-expression analysis and proceeded to subsequent analyses without removing the cancer libraries.

2.3.3. Cross-platform correlation analysis

Considering that the levels of consistency between subsets of data from a single platform were relatively low (when replicates and similar experiments were kept together) it was not surprising that datasets from different platforms compared poorly against each other. All comparisons were found to have significant but poor positive correlations when compared to randomly permuted data ($p < 0.001$, 1,000 permutations). Affymetrix versus cDNA showed the best correlation of 0.102, then Affymetrix versus SAGE with 0.086, and finally cDNA versus SAGE with 0.041 (Figure 2.9). A Pearson rank analysis also showed significant but poor agreement with only 3-8% better performance than randomly permuted data (see Figure 2.10 and section 2.3.4).

An analysis of correlation at different minimum Pearson cutoffs (r-cutoff) for gene pairs was performed as described previously [27] (Figure 2.11). Lee *et al*. (2003) observed a steady increase in global concordance (r_c = correlation of correlations) up to 0.92 at an r-cutoff of 0.91. Our data did not show such an obvious trend. Global concordance stayed close to zero (or even below) for all three pairwise platform comparisons up to 0.5-0.6 Pearson cutoff. The Affymetrix/cDNA correlation did show an improvement to $r_c = 0.163$ ($p = 0.003$, $n = 289$ gene pairs) at a r cutoff = 0.65. Similarly the Affymetrix/SAGE comparison improved to $r_c = 0.290$ (p

= 0.028, n = 44 gene pairs) at an r-cutoff = 0.7. After these cutoffs, both Affymetrix/cDNA and Affymetrix/SAGE comparisons returned to r_c values close to zero (or below) and were reduced to insignificant gene pair numbers. The cDNA/SAGE comparison showed no significant increases in r_c with any r-cutoff.

2.3.4. Ranked match analysis

The ranked match analysis showed that different expression platforms do sometimes identify the same co-expressed genes (Figure 2.10). It may be that for gene A, SAGE experiments identify its most similar gene (in terms of expression patterns) to be gene B with a Pearson correlation of 0.9. The cDNA microarray data might also identify gene B as the closest gene to A but with a Pearson value of 0.7. In such a case, a comparison of Pearson ranks may be a more useful method for evaluating cross platform consistency than actual Pearson values. The Affymetrix/cDNA comparison found that 26.5% of genes have a co-expressed gene of Pearson rank 10 or better confirmed by both platforms compared to 18.9% for random data. Affymetrix versus SAGE agreed for 26.4% of genes compared to 18.9% for random, and cDNA versus SAGE for 21.8% compared to 18.8% for random. The high percentages of genes in agreement for random data were the result of our MCE criteria. Each gene pair was required to have at least an MCE of 95, 28 or 23 (for Affymetrix, cDNA and SAGE respectively). Some genes had close to this number of experiments and thus realized the required MCE for only a few gene pair comparisons. Since we only considered gene pairs that were common in all three datasets, there were some genes that only had a little more than 10 gene pairs. In these cases, a shared match within a rank of 10 for the two platforms would occur quite commonly by chance. Thus, it is the difference over random, rather than the actual percentage, that indicates a significant number of shared matches. In all three comparisons, the percentage of shared matches observed was significantly greater than that observed between randomized datasets ($p<0.001$, 1,000 randomizations). We can conclude that the platform comparisons do identify more of the same co-expressed genes than expected by chance. However, in general the platforms show poor agreement.

2.3.5. Gene ontology analysis

Since the datasets under study demonstrated little agreement, we attempted to determine which dataset was most ‘biologically relevant’. GO biological process domain knowledge [48] was used to evaluate gene coexpression predictions for each platform. We hypothesized that genes

which are coexpressed are more likely to be involved in the same biological process. The number of gene pairs annotated to the same 'most specific' GO (Biological Process) term for each platform was determined (Figure 2.3). In general, the datasets from all platforms perform better than expected by chance. Affymetrix performed best, followed by cDNA microarray and SAGE which performed about equally better than randomly permuted data. The analysis was also extended up the GO hierarchy to parent and grandparent terms, and identical trends and relationships were observed (Figure 2.12).

A second analysis looked at the relationship between the Pearson correlation and performance against GO. For each platform, the number of gene pairs annotated to the same 'most specific term' at different Pearson correlation ranges was determined (Figure 2.4). Generally, as Pearson correlation for a gene pair increased it was more likely to be confirmed by GO. With a Pearson value in the range of 0.3-0.4 or better the platforms always performed significantly better than randomly permuted data ($p < 0.001$, 1,000 permutations). The improvement over randomly permuted data was very slight for the cDNA and SAGE datasets (2-4%). However, for the Affymetrix data, the trend was striking. Gene pairs identified as coexpressed with a Pearson correlation of 0.9-1.0 were confirmed by GO in 74% of cases. Gene pairs from this list include a large set of highly coexpressed protein biosynthesis genes as well as a few genes involved in translational elongation (a sub-process of protein biosynthesis) and muscle contraction. It should be noted that, in the case of the SAGE and cDNA datasets, only a few gene pairs had Pearson correlations > 0.9 compared to Affymetrix data (1 for cDNA, 5 for SAGE, 156 for Affymetrix).

A third analysis examined the effect of averaging platform results and comparing to individual platforms using GO. Requiring coexpression evidence from multiple datasets may represent a method of reducing noise, and increasing our confidence that coexpressed genes are actually coregulated. The percentage of gene pairs annotated to the same 'most specific term' at different average Pearson correlation ranges was determined as above. The results were again quite striking. With a 2-platform combined (2PC) Pearson of 0.4 or greater the combined platforms all performed significantly better than randomly permuted data ($p < 0.005$, 1,000 permutations). Furthermore, for any platform combination, a gene pair with an average Pearson correlation of $r > 0.6$ was much more likely to share a GO term than a gene pair with this level of correlation in only a single platform (Figure 2.5). For example, a gene pair with a two-platform average Pearson of 0.7-0.8 was found to share a common GO term 40-50% of the time. Pairs with this

same Pearson range in individual datasets shared a common GO term only 5-10% of the time, only a few percent better than expected by chance. Gene pairs confirmed by multiple datasets ($r_{avg} > 0.6$ for any two-platforms) covered a wide range of GO categories (52 in total) (Figure 2.13 and Table 2.2)

2.3.6. Comparison to other coexpression methods

Finally, an analysis was conducted to assess two other recent coexpression studies that were published while our study was in progress. The ArrayProspector (AP) method [49], the multiple microarray (TMM) method [50], and our 2-platform combination (2PC) method were each mapped to uniprot IDs and assessed using the same GO analysis as above. In all three cases, we observed significantly more gene pairs with common GO terms at higher scores (Figure 2.6). For our method (2PC), the percent of gene pairs with a common GO term rises sharply at a score of approximately 0.6-0.7. For, ArrayProspector this occurs at a score of approximately 0.7-0.8 and for TMM at a score of 5-6. At these cutoffs, each method represents 2,500 to 10,000 gene pairs. Each utilizes different genes and identifies different gene pairs as highly coexpressed. A comparison of the highest-scoring 2,500 gene pairs for each found only a minimal overlap of less than 10% (Figure 2.6D).

2.4. Discussion

We have shown that the genes identified as coexpressed are highly dependent on the dataset and expression platform used. In general, we find that the more data a correlation is based on, the more reproducible it is. When division of samples takes similar or replicate experiments into consideration, Affymetrix and cDNA internal consistencies level off at approximately $r_c = 0.25$ with MCE of about 90 and 30-40 respectively. The SAGE dataset continued to improve to nearly $r_c = 0.6$ with MCE of 80. This may reflect the diverse nature of the SAGE dataset for which libraries are rarely constructed from the same or similar tissue. In contrast, it is not uncommon for many Affymetrix or cDNA experiments to measure expression of a very similar series of samples. A recent yeast study found that the ability to correctly identify coregulated genes from coexpression analyses is highly dependent on the number of experiments with accuracy levelling off at 50 to 100 experiments [52]. Our results agree closely with this observation for human data and suggest that coexpression predictions will be most reproducible if based on at least 30 to 100 experiments. Furthermore, global coexpression analysis may benefit from a greater representation of tissues and conditions rather than greater numbers.

Given that different experimental subsets of the same platform show poor correlation it is perhaps not surprising that inter-platform comparisons show very poor correlations ($r < 0.11$). The fact that none of these data sets agree well raises some serious questions about their use for validation and integration with other data. There are several possible explanations for this observation: (1) The data comprising these datasets are so noisy as to prevent reliable identification of many truly coexpressed genes; (2) The method of identifying coexpressed genes is inadequate; (3) The unmatched and non-overlapping nature of the samples that make up each dataset result in identification of different subsets of truly coexpressed genes; and (4) Genes are under such complex regulatory control that genes coregulated in one cell-type or tissue behave in an entirely different manner in others and are therefore not globally coexpressed. It is likely that each of the explanations outlined above is to some degree responsible for the lack of concordance between coexpression analyses produced from different datasets and different platforms. It is not the purpose of this study to identify which is most important. Rather, we wish to make researchers aware that the choice of dataset or platform for integration or validation of other data could dramatically affect their results and methods that integrate or combine different platforms may be more appropriate.

It is likely that the data from each platform could be improved by altering mapping, normalization, etc. These improvements would likely increase the overall correlation. For example, a recent study showed that sequence-based (instead of annotation-based) matching between oligo and cDNA arrays significantly improved platform concordance [35, 53]. And, as mentioned above, Yauk and Berndt in their review of platform comparison studies found that correlations have steadily improved with the recent moves towards standardization and optimization of analysis methods [26]. The purpose of the current study was not to identify the best data processing methods for each platform. Instead, we attempted to choose the most standard or commonly used methods of identifying coexpressed genes and demonstrate that coexpression predicted by multiple platforms is more reliable. Future analyses that make use of coexpression could incorporate improved mapping methods, normalization, and distance metrics. Updated gene lists using our methods are provided on a separate website to include new sources of data, other species, and new mapping methods (http://www.bcgsc.ca/gc/bomge/coexpression/).

The fact that intra-platform comparisons show some correlation and improve with number of data points suggests that some gene pairs identified are truly coexpressed. Furthermore, the GO analysis shows that gene pairs identified as highly coexpressed (higher Pearson correlation) are more likely to share the same biological process and thus actually be related. Similarly, gene pairs with lower Pearson correlations were as or less likely than random chance to share the same biological process. These results suggest that the Pearson correlation is a useful metric and that both high and low Pearson values have the meaning we expect. The GO analysis did not conclusively identify a single 'correct' platform or dataset but it did show that the Affymetrix dataset identified more biologically relevant gene pairs than the cDNA or SAGE datasets. However, gene pairs coexpressed in multiple expression platforms were much more likely to be confirmed by GO. Thus, combining platforms appears to act as a filter, producing high-confidence predictions from noisy datasets. This conclusion is based on the assumption that coexpressed genes are more likely to be biologically relevant if they share common biological processes. Assessments using GO are limited by issues such as the incompleteness of the ontology, the potential for circularity (addressed in Methods), experimental bias towards 'well-studied' genes, and inconsistencies in structure and depth. Furthermore, it is likely that some coregulated genes will belong to different biological processes while other genes involved in the same process will not be coregulated. As such, an 'absolute' performance against GO is difficult or impossible to define. Despite these issues, we believe GO currently represents one of the best resources for a relative assessment of coexpression platforms or methods. In a recent systematic survey of the "guilt-by-association" (GBA) method, Wolfe *et al.* (2005) investigated the functional organization of five coexpression networks from three mammalian organisms. They found that the signature of GBA is ubiquitous and reproducible and that the GBA method is broadly applicable across entire the Gene Ontology [33]. To put it in simple terms, gene coexpression and GO coannotation are closely related. Genes that are coexpressed are more likely to share common biological processes and vice versa, and this pattern holds true across multiple species and expression datasets.

Recent investigations into the utility of combining expression data from different high-throughput platforms have identified highly variable levels of agreement. Based on an analysis of a small set of matched samples using oligonucleotide arrays, SAGE, and EST data, Haverty and colleagues [54] caution against the combination of platforms to confirm expression patterns for specific sets of genes. However, they do suggest that such methods can be used to extract

high-confidence subsets of related genes. We agree that for many genes a poor level of agreement between datasets raises questions about their utility. However, our results do show that platform combination methods can be extended to large sets of unmatched publicly available expression data to produce biologically meaningful information.

As we were nearing completion of our analysis, a similar study using multiple microarray datasets (TMM, the multiple microarray dataset) was published. Lee *et al.* (2004) examined 60 microarray datasets (cDNA and oligonucleotide) for gene pairs identified as coexpressed in multiple datasets [50]. They report that even gene pairs confirmed by only a single dataset have better GO similarity scores than random pairs and GO score increases steadily with the number of confirmed links. Their method differs from ours in that experimental subsets are analyzed separately and a 'vote-counting' method was used to identify gene pairs that appear highly coexpressed (above some Pearson cutoff) in multiple sets. Our method combines all experimental subsets into a single dataset for each expression platform and then averages the global Pearson correlations between platforms. Our method is also the first to include SAGE data. A third recently published method (ArrayProspector), used a combination of singular value decomposition and kernel density estimation [49]. This method combines evidence from related arrays and weights the contribution of each array according to how well they correlate with functional annotation.

When attempting to infer function or coregulation from coexpression we should consider that it is likely that genes are biologically related in a number of different ways and therefore different methods will be required to identify each type of relationship. For example, one pair of genes might be 'tightly' coexpressed only under very specific conditions whereas another gene pair might be 'loosely' coexpressed across a broad range of conditions depending on the regulatory elements that they share. The three methods discussed above (TMM, AP, and 2PC) represent three different approaches to the problem of identifying high-confidence coexpression for the purpose of inferring function or coregulation. Because the methods use different datasets, scoring methods, and comprise different gene sets, a direct comparison of the methods is difficult. Therefore, we chose to simply assess their respective predictions against GO independently. Thus, we do not identify the 'best' method but rather show that each method is at least partially effective based on performance against the Gene Ontology. Furthermore, because the highest-scoring pairs for each are almost completely non-overlapping we advocate

combining the best results of each into a single set of high-confidence predictions. To this end we have chosen score thresholds for each method based on GO performance (2PC > 0.65; AP > 0.7; TMM > 7) and make available a list of 13,145 high-confidence coexpressed gene pairs (representing 2,979 unique genes) (http://www.bcgsc.ca/gc/bomge/coexpression/suppl_materials) for use in regulatory element prediction or other integration studies.

2.5. Conclusions

The analysis presented in this chapter opened three avenues for further inquiry. First, we found that in general, global coexpression measurements are not very reproducible between different platforms. One possible explanation for this (discussed above) is that few genes are actually globally coregulated across most or many different tissues and conditions. Instead, it may be that most genes are coregulated in different combinations under more specific conditions. To investigate this possibility, "biclustering" or "subspace clustering" methods are needed which identify potentially overlapping groups of genes that are tightly coexpressed only under specific subsets of the samples assayed. In chapter three, we describe the development and assessment of such a method that is capable of processing large expression datasets like those presented in this chapter. Secondly, we showed that combining coexpression measurements from multiple platforms increased confidence in coexpression predictions. In chapter four we investigate whether this multi-platform strategy can be applied to differential expression instead of coexpression. Finally, we identified high-confidence coexpressed genes for the ultimate purpose of identifying coregulated genes and their regulatory element sequences. However, a major barrier to identifying regulatory elements is the lack of positive and negative control sequences (i.e., experimentally proven regulatory sequences or sequences verified to have no regulatory function). In chapter five we present a database and web resource called ORegAnno to address this challenge.

Table 2.1 Summary of r_c values for internal consistency analysis using different sample division methods and MCE cutoffs

Note that many different divisions are possible for each result below. Gene pair and r_c values represent mean values from 100 different random or pseudo-random divisions.

Platform	Division	MCE cutoff	Gene pairs	r_c value
Affymetrix	Random	100	4,149,092	0.925
	Pseudo-random by GSE series	95	3,427,174	0.257
		100	3,260,557	0.253
cDNA Microarray	Random	100	10,429,219	0.889
	Pseudo-random by author	28	11,178,346	0.253
		100	9,747,169	0.273
SAGE	Random	100	2,635	0.776
	Pseudo-random by tissue	23	577,820	0.253
		100	1,518	0.660

Table 2.2. GO categories for gene pairs confirmed by 2-platform combination (2PC) method

Gene Pairs	Percent	Common Term	Go Term
257	55.508	GO:0006412	protein biosynthesis
25	5.3996	GO:0006355	regulation of transcription, DNA-dependent
18	3.8877	GO:0007067	mitosis
15	3.2397	GO:0006260	DNA replication
14	3.0238	GO:0006281	DNA repair
13	2.8078	GO:0006468	protein amino acid phosphorylation
11	2.3758	GO:0006958	complement activation, classical pathway
10	2.1598	GO:0008152	metabolism
8	1.7279	GO:0007049	cell cycle
7	1.5119	GO:0000910	cytokinesis
6	1.2959	GO:0006810	transport
6	1.2959	GO:0007165	signal transduction
5	1.0799	GO:0000074	regulation of cell cycle
5	1.0799	GO:0006118	electron transport
4	0.8639	GO:0008283	cell proliferation
4	0.8639	GO:0006955	immune response
3	0.6479	GO:0006414	translational elongation
3	0.6479	GO:0000398	nuclear mRNA splicing, via spliceosome
3	0.6479	GO:0006357	regulation of transcription from Pol II promoter
3	0.6479	GO:0006936	muscle contraction
2	0.432	GO:0007155	cell adhesion
2	0.432	GO:0006508	proteolysis and peptidolysis
2	0.432	GO:0006464	protein modification
2	0.432	GO:0007517	muscle development
2	0.432	GO:0007275	development
2	0.432	GO:0006928	cell motility
2	0.432	GO:0006511	ubiquitin-dependent protein catabolism
2	0.432	GO:0006350	transcription
2	0.432	GO:0008151	cell growth and/or maintenance
2	0.432	GO:0007264	small GTPase mediated signal transduction
2	0.432	GO:0006418	tRNA aminoacylation for protein translation
2	0.432	GO:0006915	apoptosis
1	0.216	GO:0015031	protein transport
1	0.216	GO:0006461	protein complex assembly
1	0.216	GO:0008380	RNA splicing
1	0.216	GO:0006364	rRNA processing
1	0.216	GO:0006366	transcription from Pol II promoter
1	0.216	GO:0007242	intracellular signalling cascade
1	0.216	GO:0006091	energy pathways
1	0.216	GO:0006635	fatty acid beta-oxidation
1	0.216	GO:0006177	GMP biosynthesis
1	0.216	GO:0006096	glycolysis
1	0.216	GO:0006259	DNA metabolism
1	0.216	GO:0007186	G-protein coupled receptor protein signalling pathway
1	0.216	GO:0007267	cell-cell signalling
1	0.216	GO:0006098	pentose-phosphate shunt
1	0.216	GO:0006917	induction of apoptosis
1	0.216	GO:0016575	histone deacetylation
1	0.216	GO:0006954	inflammatory response
1	0.216	GO:0045786	negative regulation of cell cycle
1	0.216	GO:0009966	regulation of signal transduction
463	**100**	**52**	

Figure 2.1. Venn diagram outlining datasets used in analysis

N indicates the number of experiments available for the platform. For Affymetrix, the number in brackets indicates the subset of experiments providing detection (present (P); marginal (M); absent (A)) calls. The number of genes represents only those genes that could be unambiguously mapped to a LocusLink ID.

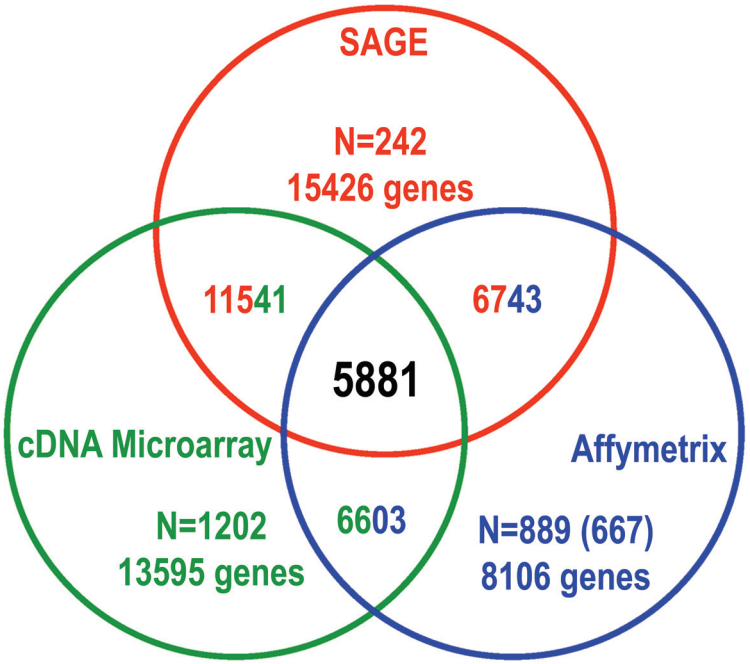


Figure 2.2. Internal consistency and minimum common experiments analysis using pseudo-random division method

A-C. For each gene pair, the number of common experiments was determined as the number of experiments for which expression values were available for both genes. On the left axis, minimum common experiments (MCE) is plotted against internal consistency. On the right axis, MCE is plotted against number of gene pairs. Data represent mean r_c value and gene pair number of 100 pseudo-random divisions at each MCE. Error bars indicate one standard deviation. D. A schematic explaining the concept of MCE is shown. In this imaginary dataset, there are three genes (GeneA, GeneB, and GeneC) with some expression value for five experiments (Exp1 through Exp5). GeneA and GeneB both have values for Exp1, Exp4, and Exp5. A measure of their coexpression (e.g. Pearson correlation) could be calculated using data from these three common experiments only. If an MCE threshold of 3 was desired, the gene pair for GeneA/Gene B would meet the threshold but the other gene pairs (GeneA/GeneC and GeneB/GeneC) would not.

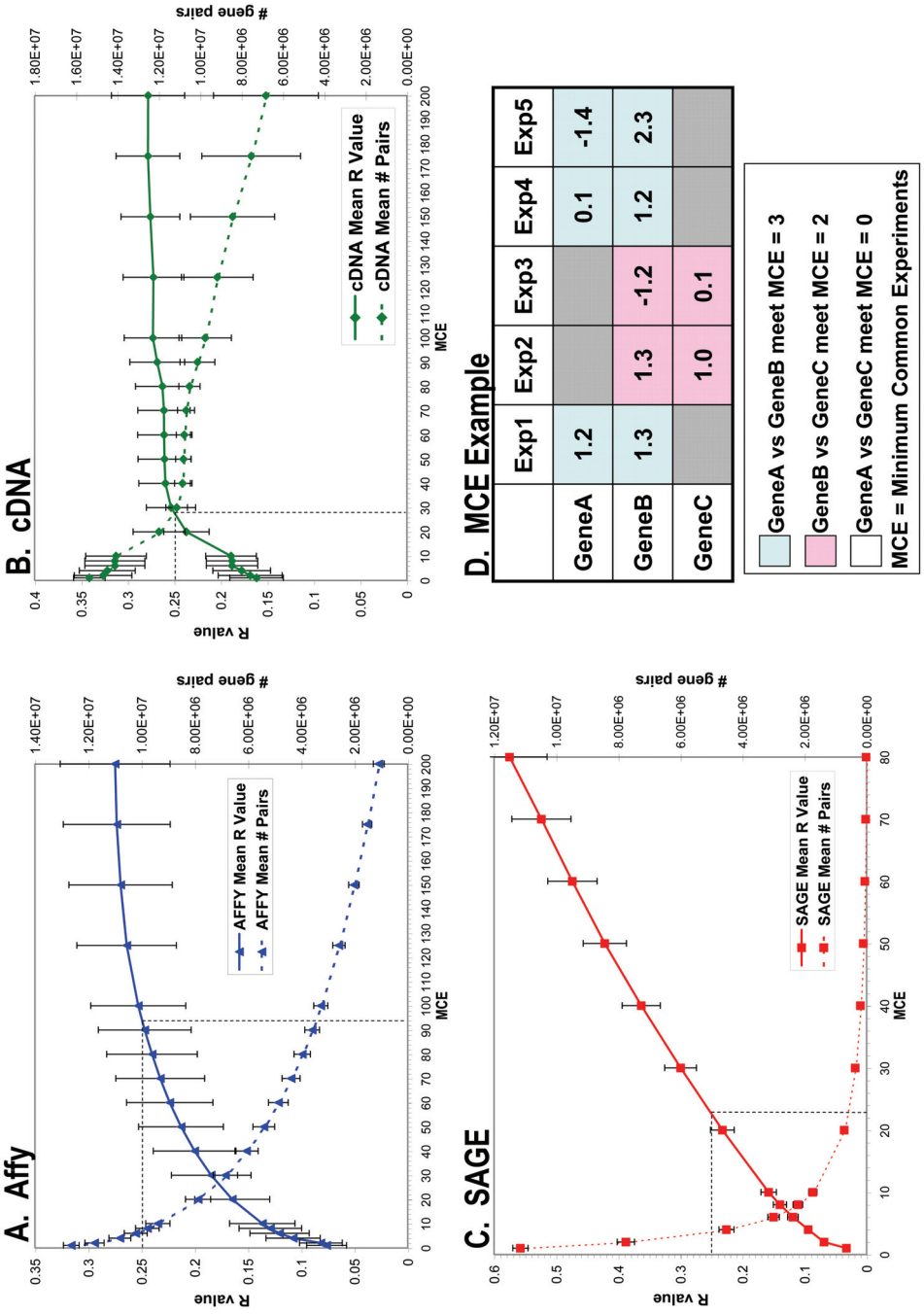


Figure 2.3. GO analysis

Gene pairs for which both genes were annotated with Gene Ontology Biological Process terms were evaluated to determine the percentage of pairs within a neighbourhood of k that are annotated with the same GO term. As the GO annotation is hierarchical, only the most specific GO terms for each gene were considered. Comparison of these percentages to results produced from randomizing gene pair correlations indicate that gene pairs found to be correlated by any platform were more likely to share the same function than randomly chosen gene pairs ($p<0.001$, 1,000 randomizations). Affymetrix appears to predict the most biologically relevant gene pair correlations.

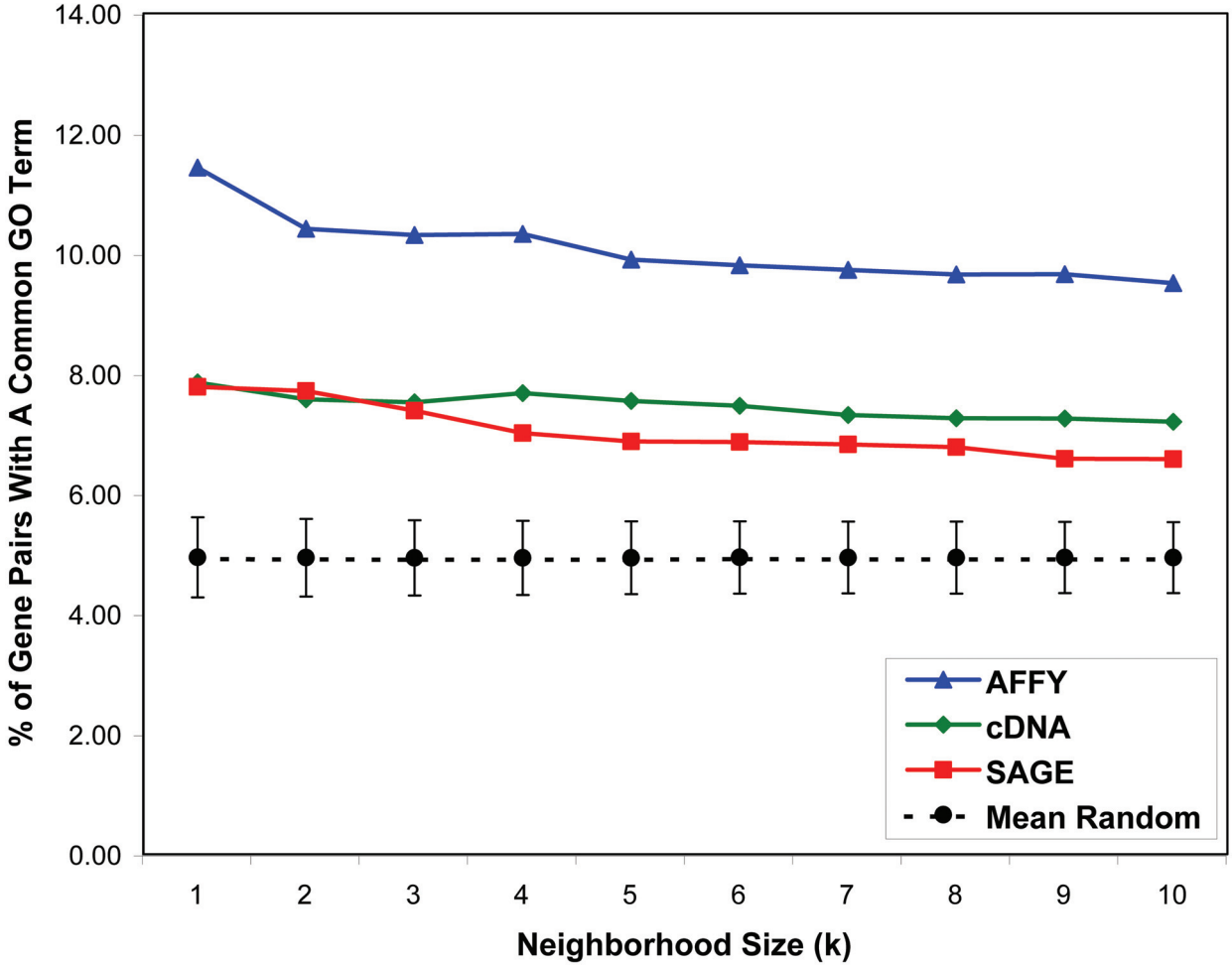


Figure 2.4. GO correlation range analysis

For each platform, the number of gene pairs annotated to the same 'most specific term' at different Pearson correlation ranges was determined. Generally, as Pearson correlation for a gene pair increased it was more likely to be confirmed by GO. With a Pearson value in the range of 0.3-0.4 or higher the platforms always performed significantly better than randomly permuted data ($p < 0.001$, 1,000 permutations). The improvement over randomly permuted data was very slight for the cDNA and SAGE datasets (2-4%). However, for the Affymetrix data, the trend was striking. Gene pairs identified as coexpressed with a Pearson correlation of 0.9-1.0 were confirmed by GO in 74% of cases. Random lines represent mean values from 1,000 random permutations. Error bars indicate one standard deviation. Inset graphs represent the same data but with y-axis scale showing only the bottom 10%.

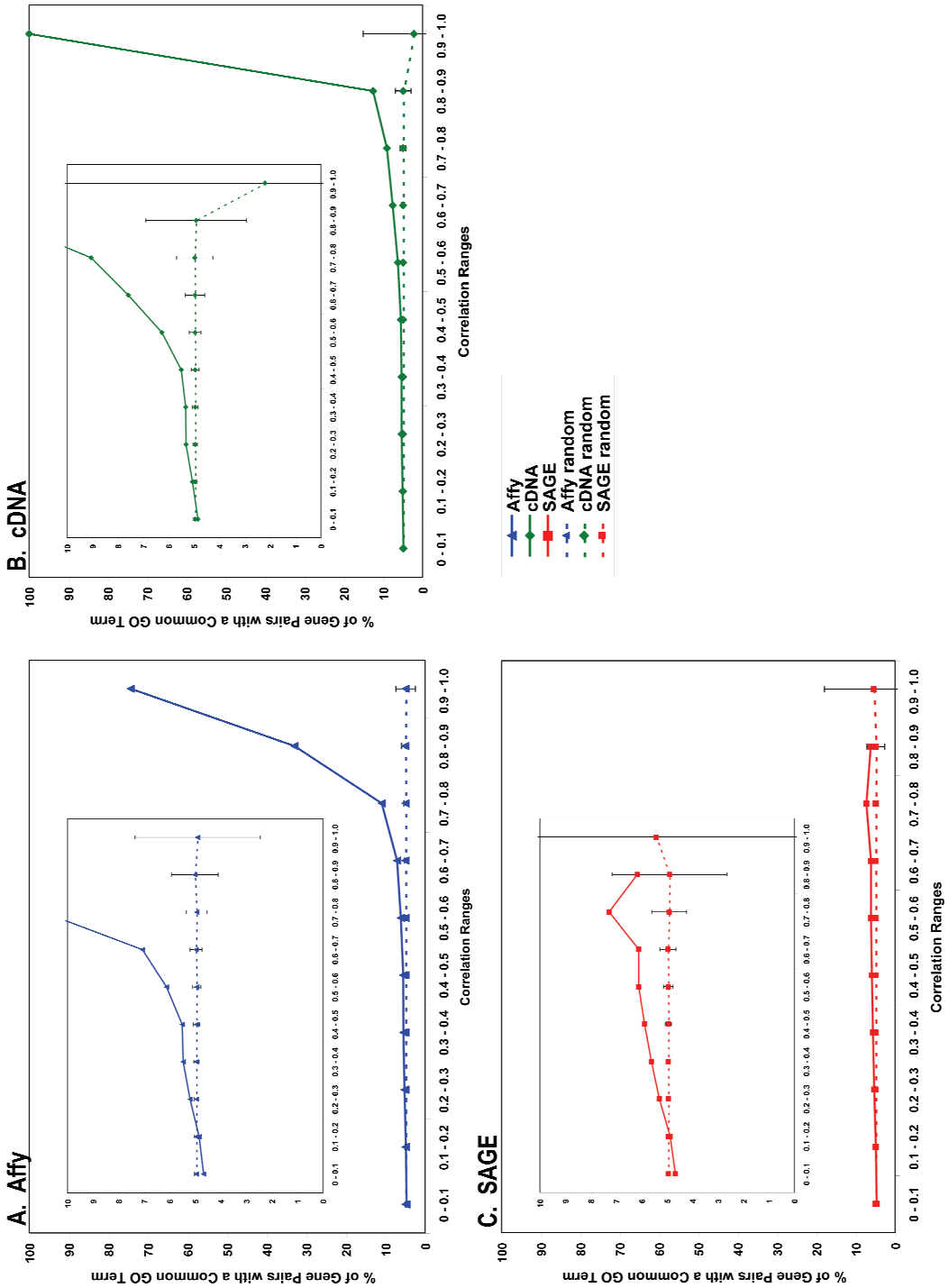


Figure 2.5. GO correlation range analysis for multi-platform average

The effect of averaging coexpression between platforms is shown. Requiring coexpression evidence from multiple datasets may represent a method of reducing noise and increasing our confidence that coexpressed genes are actually coregulated. The percentage of gene pairs annotated to the same 'most specific term' at different average Pearson correlation ranges was determined. With a 2-platform combined (2PC) Pearson of 0.4 or greater the combined platforms all performed significantly better than randomly permuted data ($p < 0.005$, 1,000 permutations). Furthermore, for any platform combination, a gene pair with an average Pearson correlation of $r > 0.6$ was much more likely to share a GO term than a gene pair with this level of correlation in only a single platform. For example, a gene pair with a two-platform average Pearson of 0.7-0.8 was found to share a common GO term 40-50% of the time. Pairs with this same Pearson range in individual datasets shared a common GO term only 5-10% of the time. Random line represents mean values from 1,000 random permutations of all two-platform combinations. Error bars indicate one standard deviation.

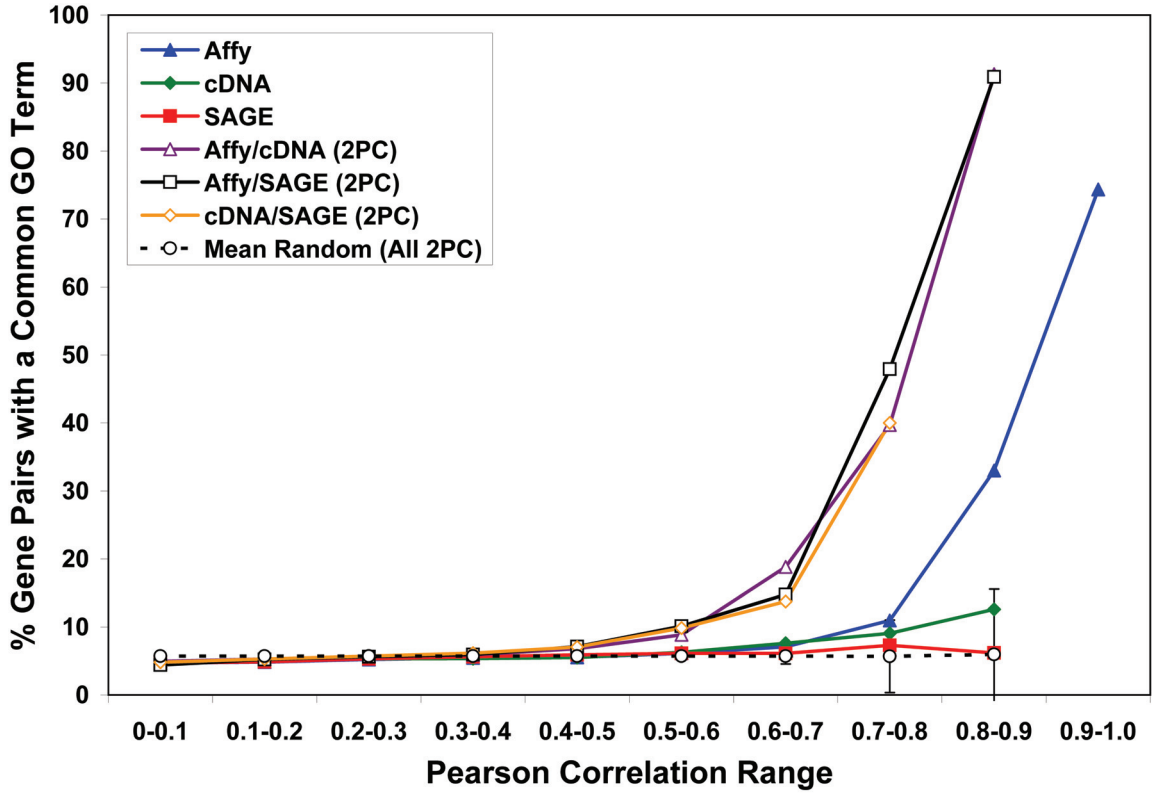


Figure 2.6. Comparison of 2-platform combination method to other recent coexpression methods

A-C. An analysis was conducted to assess two other recent coexpression studies. The ArrayProspector (AP) method [49], the multiple microarray (TMM) method [50], and our 2-platform combination (2PC) method were each mapped to uniprot IDs and assessed using the same GO analysis. In all three cases, we observed significantly more gene pairs with common GO terms at higher scores. For our method (2PC), the percent of gene pairs with a common GO term rises sharply at a score of approximately 0.6-0.7. For ArrayProspector this occurs at a score of approximately 0.7-0.8 and for TMM at a score of 5-6. At these cutoffs, each method represents ~2,500 to ~10,000 gene pairs. Lines with hollow squares represent numbers of gene pairs (right axis). Lines with solid triangles represent % of gene pairs with a common GO term (left axis). Random lines represent mean values from 1,000 random permutations. Error bars indicate one standard deviation. The legend for all three panels (A-C) is shown in panel C only. D. The Venn diagram indicates overlap between the 2,500 top scoring pairs for each method (not required to be in GO).

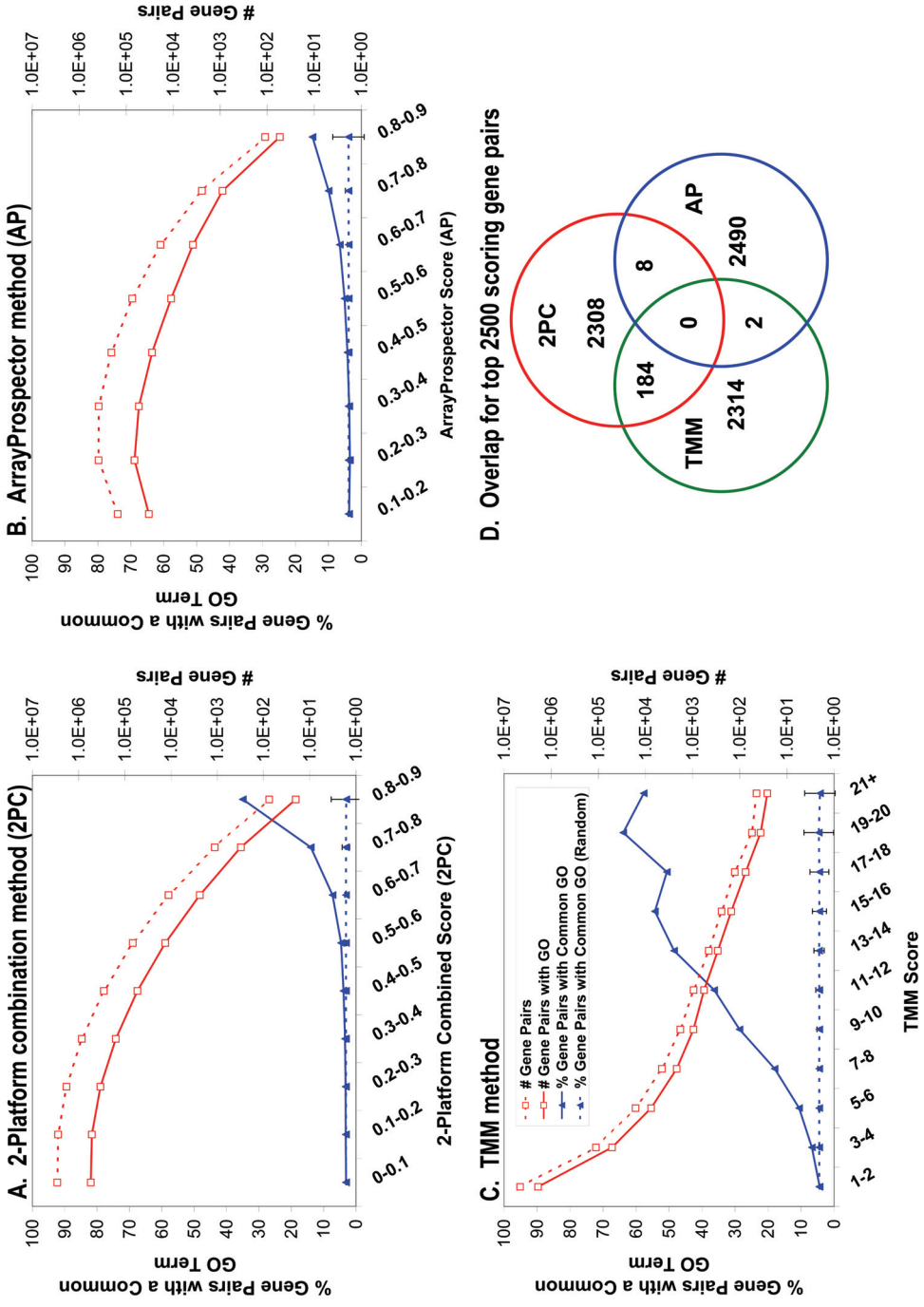


Figure 2.7. Internal consistency analysis based on random division of experiments

Analysis is identical to Figure 2.2, except division of libraries is random rather than pseudo-random (by experiment, author, or tissue), resulting in much higher r_c values due to presence of replicates or very similar experiments. Data represent mean r_c value and gene pair number of 100 random divisions. Error bars indicate one standard deviation.

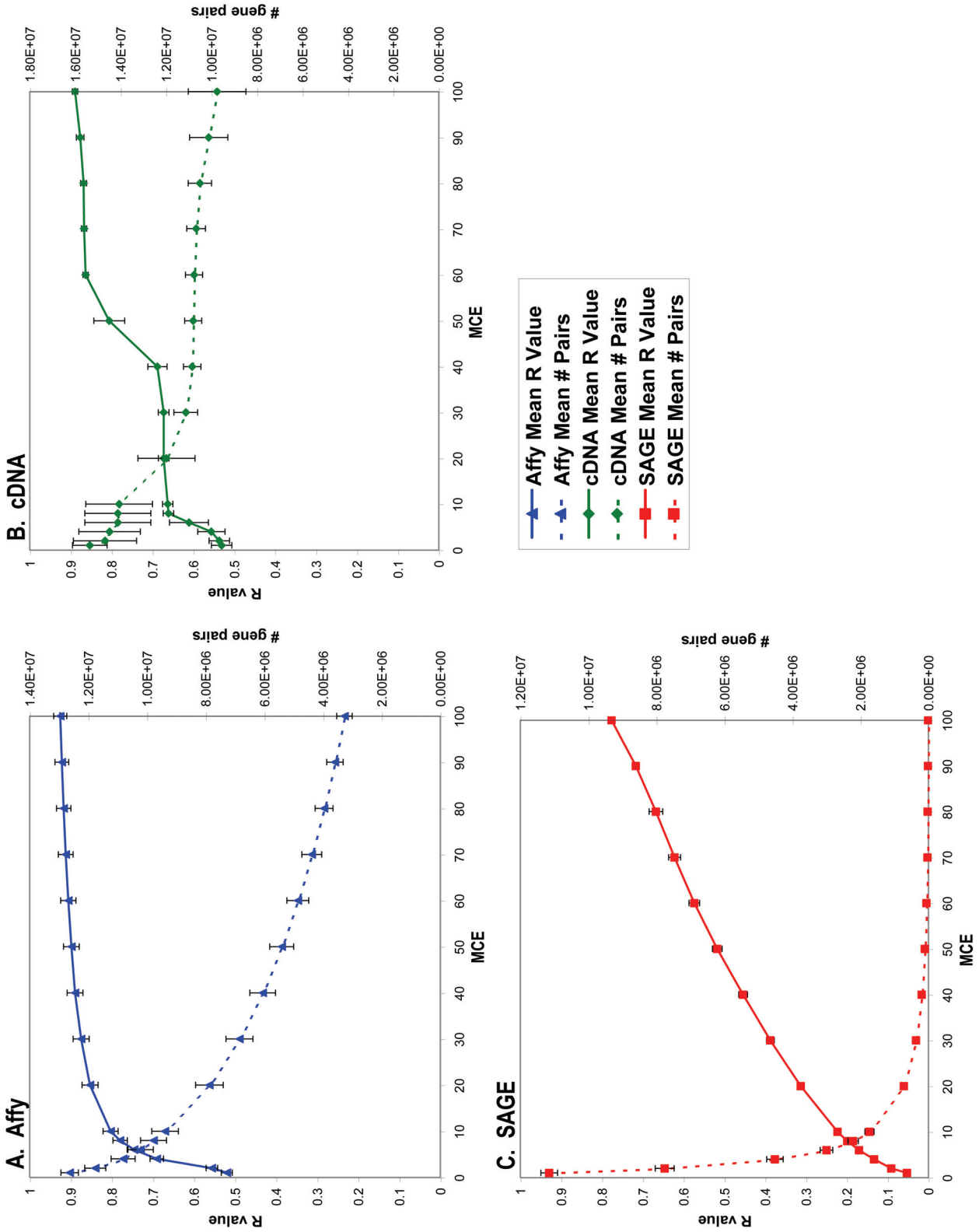


Figure 2.8. SAGE cancer versus normal analysis

Plots represent correlation of correlations for subsets of SAGE data. (A) Correlation between normal and cancer SAGE libraries, r_c=0.324 for 23 MCE; (B) Correlation between randomly divided subsets of SAGE data, r_c=0.267 for MCE of 23. The results for cancer versus normal were comparable to that seen for the internal consistency of SAGE when not taking cancer status into account.

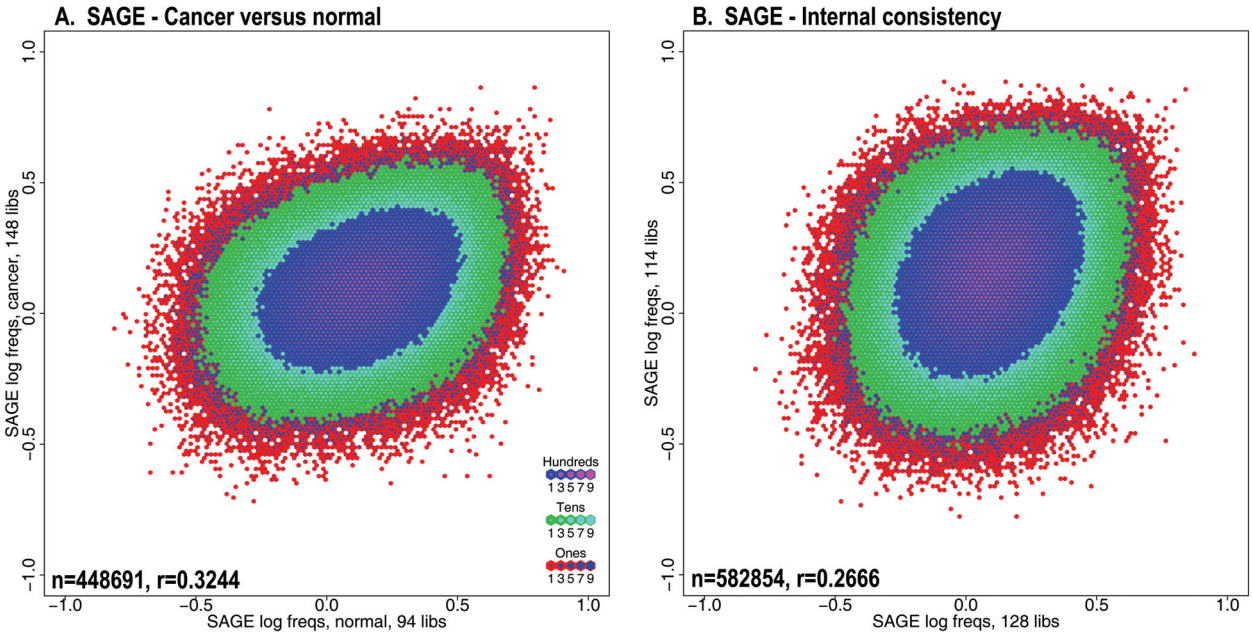


Figure 2.9. Platform comparisons

Plots represent correlation of correlations (r_c) between each pairwise platform comparison. A. Affymetrix versus cDNA, r_c=0.102; B. Affymetrix versus SAGE, r_c=0.086; C. cDNA versus SAGE, r_c=0.041. 1,173,330 gene pairs are shown representing the intersection between Affymetrix, cDNA, and SAGE for which 95, 28, and 23 MCE were required respectively for each Pearson correlation calculation. Correlations observed in A-C were significant when compared to randomized data ($p<0.001$, 1,000 randomizations) but generally poor. Small inset boxes show representative randomized data; D-F. Pearson correlation (r) frequency distributions for each platform. Notice that each displays a similar, approximately normal distribution with a slight skew towards positive correlations.

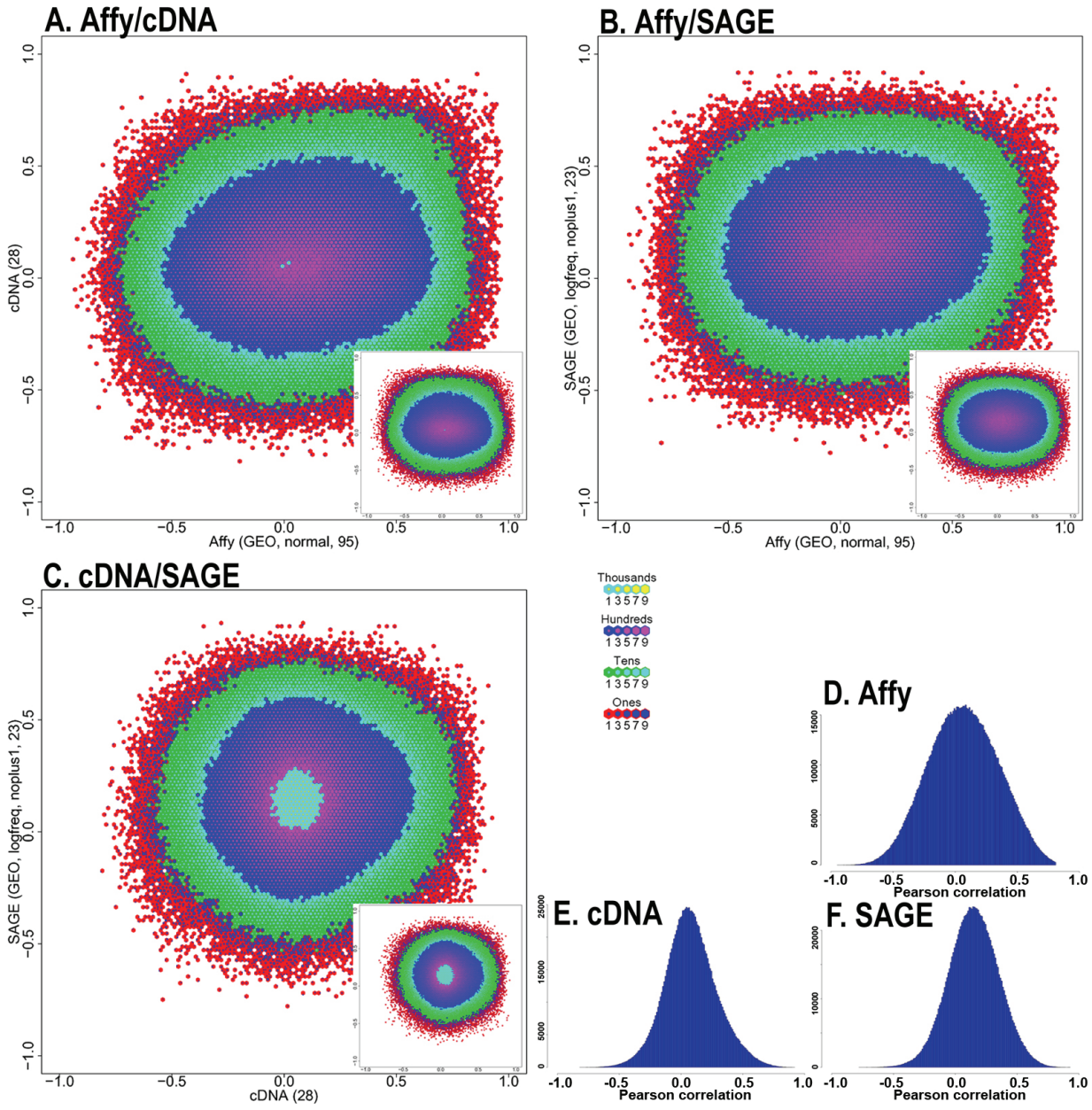


Figure 2.10. Ranked Pearson analysis

The percentage of genes with a co-expressed gene identified by both platforms within a rank or neighbourhood of k for each platform comparison is shown. The platforms identified more of the same co-expressed genes than expected by chance. However, in general the platforms showed poor agreement. Random lines represent mean values from 1,000 randomizations. Error bars indicate one standard deviation.

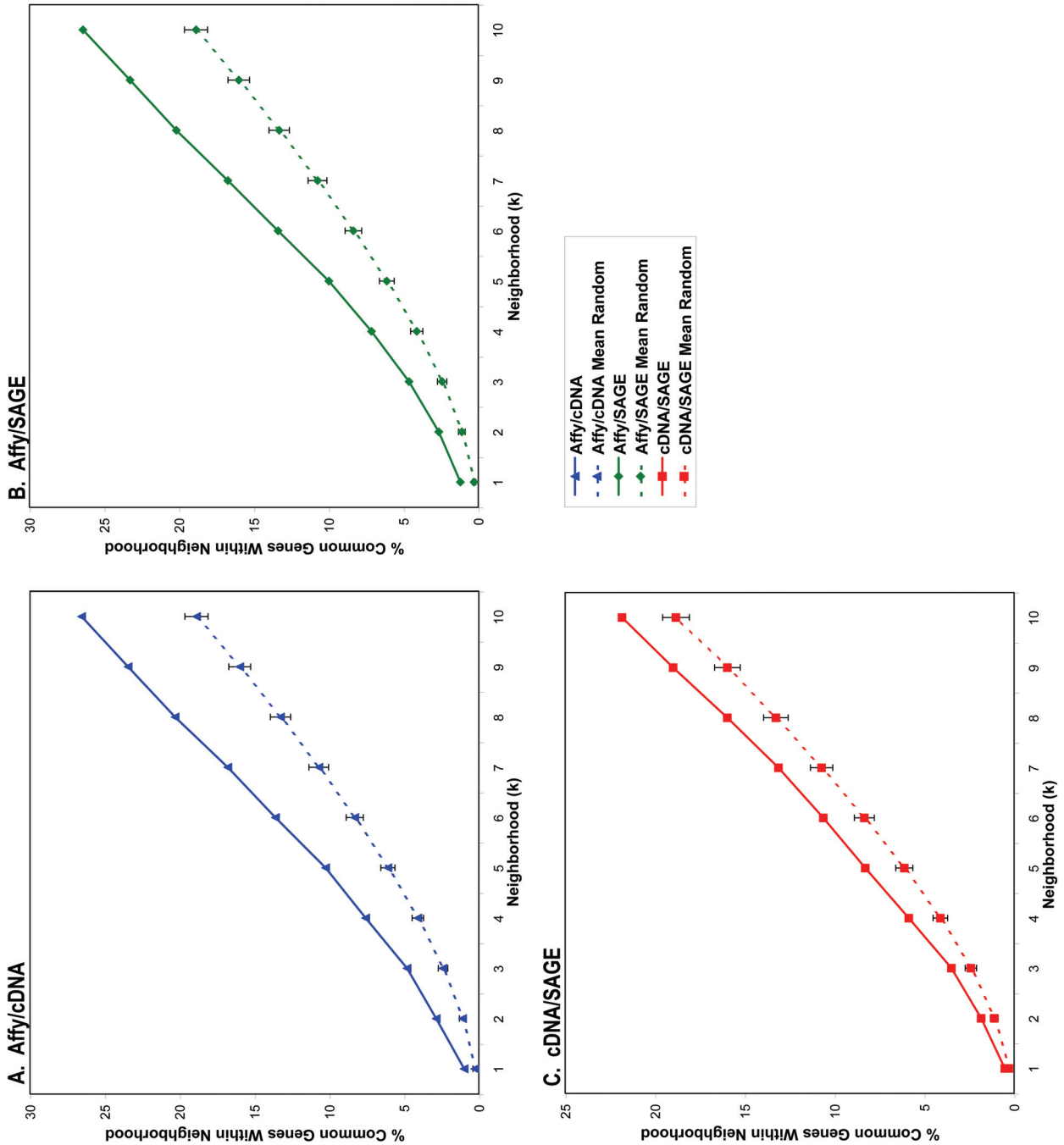


Figure 2.11. Effect of correlation cutoff on r_c

Platform comparisons were repeated with subsets of gene pairs having correlations above certain cutoffs (0.1 increments). Only positive correlations were considered. Higher global concordance was observed for the Affymetrix/cDNA comparison at a Pearson cutoff (r-cutoff) of 0.65 and for the Affymetrix/SAGE comparison at r-cutoff of 0.6 and 0.7 ($p<0.05$). The cDNA/SAGE comparison did not show any increase that was significant. In any case, the steady trend of increasing r_c with more stringent r-cutoff was not observed as reported elsewhere [27]. Asterisks indicate increased r_c values which were found to be significant ($p<0.05$).

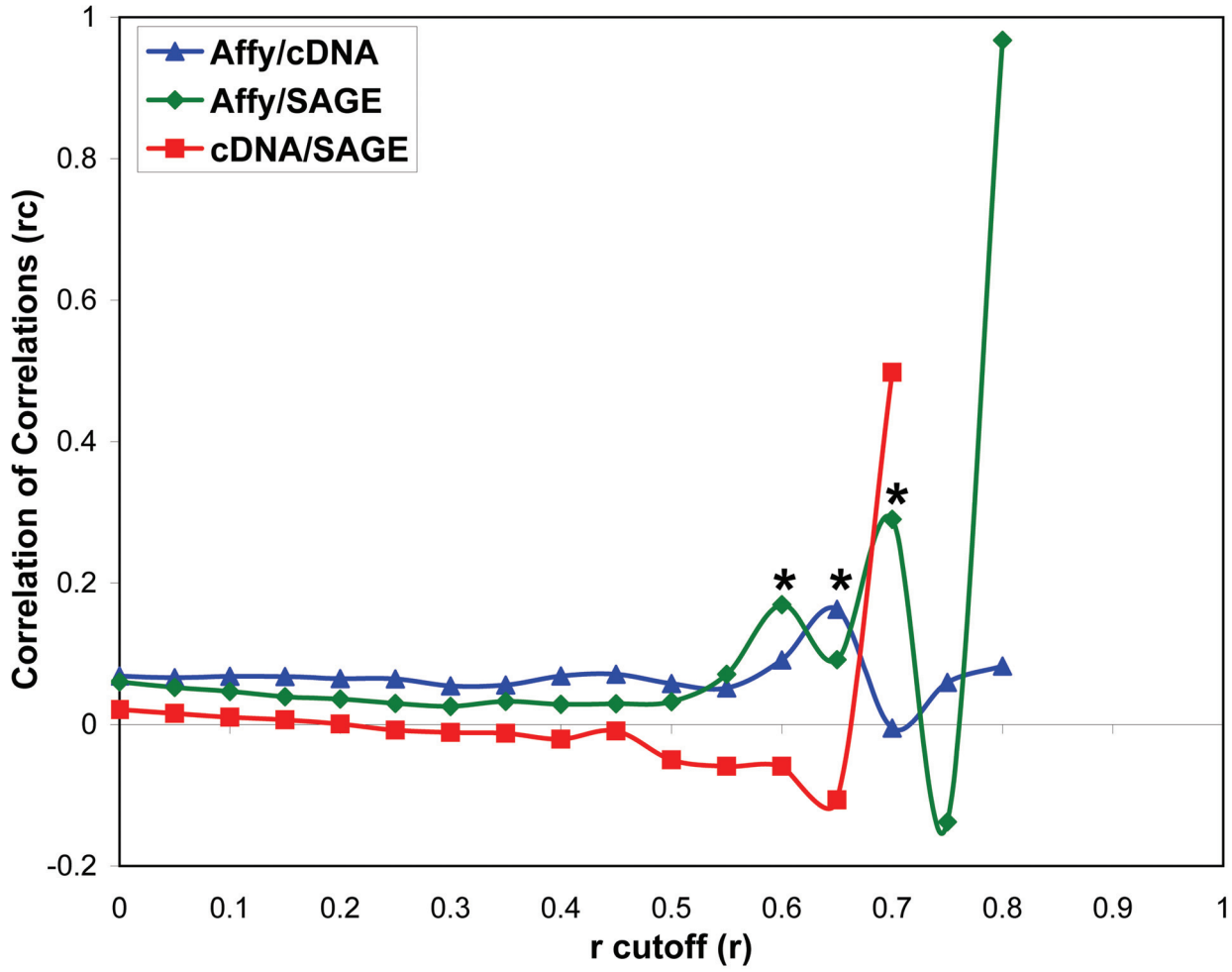


Figure 2.12. Expanded GO analysis including hierarchical relationships

In this figure, analysis was performed as for Figure 2.3, but in addition to considering only the most specific GO term annotations (A), the percentage of gene pairs sharing parent terms (B) or parent and grandparent terms (C) were also determined. As before, gene pairs found to be correlated by any platform were more likely to share the same function than randomly chosen gene pairs ($p<0.001$, 1,000 randomizations). As higher levels in the GO hierarchical tree (parent and grandparent terms) were considered, there was a higher chance that randomly chosen gene pairs would share GO terms, resulting in less difference between random and actual data. The legend for all three panels (A-C) is shown in panel A.

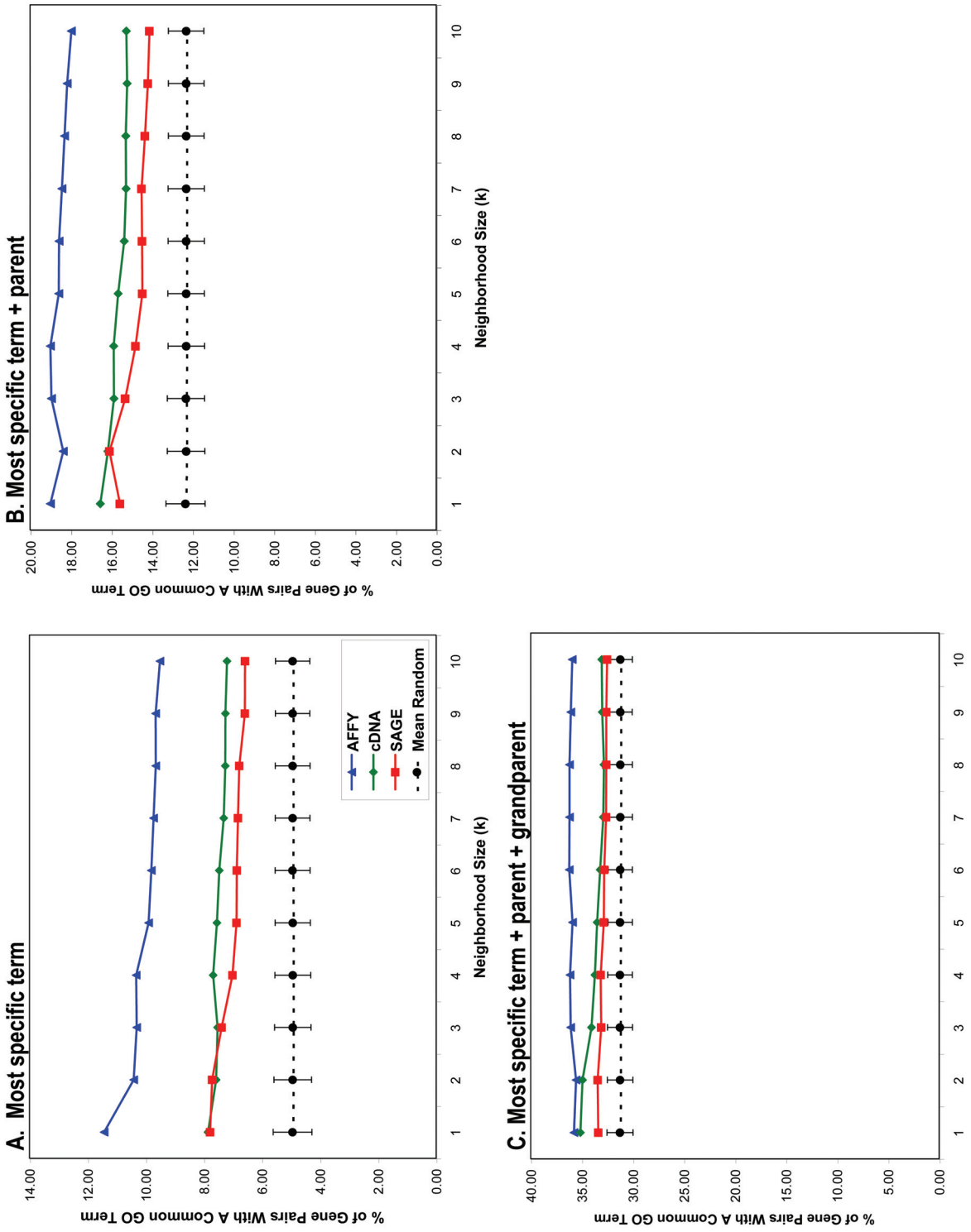
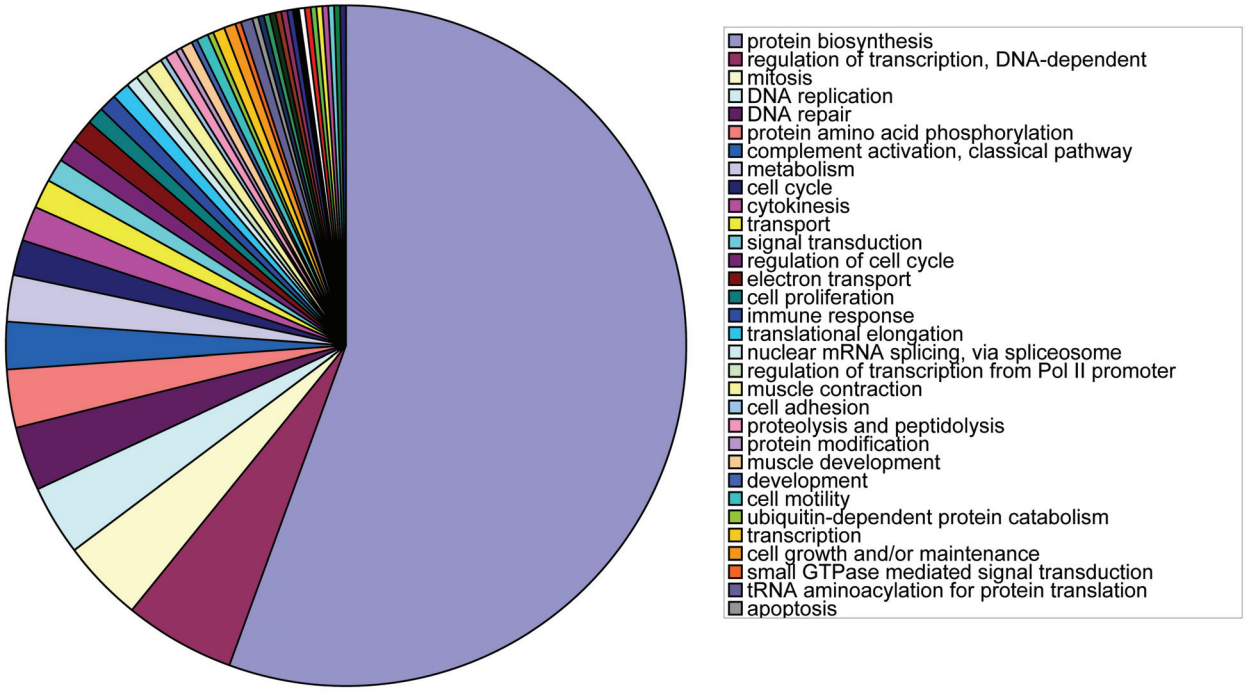


Figure 2.13. GO categories for gene pairs confirmed by multiple datasets

The chart shows GO terms of gene pairs with an average Pearson correlation of r>0.6 for any two of three platform datasets (Affymetrix, cDNA microarray, SAGE). The legend only shows the 32 categories with more than one gene pair. However, another 20 categories are represented on the chart and are summarized in Table 2.2.



References

1. Schena, M., D. Shalon, R.W. Davis, and P.O. Brown, *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
2. Lockhart, D.J., H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown, *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.
3. Velculescu, V.E., L. Zhang, B. Vogelstein, and K.W. Kinzler, *Serial analysis of gene expression.* Science, 1995. **270**(5235): p. 484-7.
4. Stuart, J.M., E. Segal, D. Koller, and S.K. Kim, *A gene-coexpression network for global discovery of conserved genetic modules.* Science, 2003. **302**(5643): p. 249-55.
5. Li, S., C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.O. Vidalain, J.D. Han, A. Chesneau, T. Hao, D.S. Goldberg, N. Li, M. Martinez, J.F. Rual, P. Lamesch, L. Xu, M. Tewari, S.L. Wong, L.V. Zhang, G.F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H.W. Gabel, A. Elewa, B. Baumgartner, D.J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S.E. Mango, W.M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K.C. Gunsalus, J.W. Harper, M.E. Cusick, F.P. Roth, D.E. Hill, and M. Vidal, *A map of the interactome network of the metazoan C. elegans.* Science, 2004. **303**(5657): p. 540-3.
6. Kemmeren, P., N.L. van Berkum, J. Vilo, T. Bijma, R. Donders, A. Brazma, and F.C. Holstege, *Protein interaction verification and functional annotation by integrated analysis of genome-scale data.* Mol Cell, 2002. **9**(5): p. 1133-43.
7. Walhout, A.J., J. Reboul, O. Shtanko, N. Bertin, P. Vaglio, H. Ge, H. Lee, L. Doucette-Stamm, K.C. Gunsalus, A.J. Schetter, D.G. Morton, K.J. Kemphues, V. Reinke, S.K. Kim, F. Piano, and M. Vidal, *Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline.* Curr Biol, 2002. **12**(22): p. 1952-8.
8. Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet, 2003. **34**(2): p. 166-76.
9. Yeoh, E.J., M.E. Ross, S.A. Shurtleff, W.K. Williams, D. Patel, R. Mahfouz, F.G. Behm, S.C. Raimondi, M.V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.H. Pui, W.E. Evans, C. Naeve, L. Wong, and J.R. Downing, *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.* Cancer Cell, 2002. **1**(2): p. 133-43.
10. Nimgaonkar, A., D. Sanoudou, A.J. Butte, J.N. Haslett, L.M. Kunkel, A.H. Beggs, and I.S. Kohane, *Reproducibility of gene expression across generations of Affymetrix microarrays.* BMC Bioinformatics, 2003. **4**(1): p. 27.
11. Tan, P.K., T.J. Downey, E.L. Spitznagel, Jr., P. Xu, D. Fu, D.S. Dimitrov, R.A. Lempicki, B.M. Raaka, and M.C. Cam, *Evaluation of gene expression measurements from commercial microarray platforms.* Nucleic Acids Res, 2003. **31**(19): p. 5676-84.
12. Dinel, S., C. Bolduc, P. Belleau, A. Boivin, M. Yoshioka, E. Calvo, B. Piedboeuf, E.E. Snyder, F. Labrie, and J. St-Amand, *Reproducibility, bioinformatic analysis and power of the SAGE method to evaluate changes in transcriptome.* Nucleic Acids Res, 2005. **33**(3): p. e26.
13. Wang, S.M., *Understanding SAGE data.* Trends Genet, 2007. **23**(1): p. 42-50.
14. Huminiecki, L., A.T. Lloyd, and K.H. Wolfe, *Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases.* BMC Genomics, 2003. **4**(1): p. 31.

15. Detours, V., J.E. Dumont, H. Bersini, and C. Maenhaut, *Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets.* FEBS Lett, 2003. **546**(1): p. 98-102.
16. Jarvinen, A.K., S. Hautaniemi, H. Edgren, P. Auvinen, J. Saarela, O.P. Kallioniemi, and O. Monni, *Are data from different gene expression microarray platforms comparable?* Genomics, 2004. **83**(6): p. 1164-8.
17. Iacobuzio-Donahue, C.A., R. Ashfaq, A. Maitra, N.V. Adsay, G.L. Shen-Ong, K. Berg, M.A. Hollingsworth, J.L. Cameron, C.J. Yeo, S.E. Kern, M. Goggins, and R.H. Hruban, *Highly expressed genes in pancreatic ductal adenocarcinomas: a comprehensive characterization and comparison of the transcription profiles obtained from three major technologies.* Cancer Res, 2003. **63**(24): p. 8614-22.
18. Kim, H.L., *Comparison of oligonucleotide-microarray and serial analysis of gene expression (SAGE) in transcript profiling analysis of megakaryocytes derived from CD34+ cells.* Exp Mol Med, 2003. **35**(5): p. 460-6.
19. Rogojina, A.T., W.E. Orr, B.K. Song, and E.E. Geisert, Jr., *Comparing the use of Affymetrix to spotted oligonucleotide microarrays using two retinal pigment epithelium cell lines.* Mol Vis, 2003. **9**: p. 482-96.
20. Ishii, M., S. Hashimoto, S. Tsutsumi, Y. Wada, K. Matsushima, T. Kodama, and H. Aburatani, *Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis.* Genomics, 2000. **68**(2): p. 136-43.
21. Evans, S.J., N.A. Datson, M. Kabbaj, R.C. Thompson, E. Vreugdenhil, E.R. De Kloet, S.J. Watson, and H. Akil, *Evaluation of Affymetrix Gene Chip sensitivity in rat hippocampal tissue using SAGE analysis. Serial Analysis of Gene Expression.* Eur J Neurosci, 2002. **16**(3): p. 409-13.
22. Li, J., M. Pankratz, and J.A. Johnson, *Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays.* Toxicol Sci, 2002. **69**(2): p. 383-90.
23. Kuo, W.P., T.K. Jenssen, A.J. Butte, L. Ohno-Machado, and I.S. Kohane, *Analysis of matched mRNA measurements from two different microarray technologies.* Bioinformatics, 2002. **18**(3): p. 405-12.
24. Mah, N., A. Thelin, T. Lu, S. Nikolaus, T. Kuhbacher, Y. Gurbuz, H. Eickhoff, G. Kloppel, H. Lehrach, B. Mellgard, C.M. Costello, and S. Schreiber, *A comparison of oligonucleotide and cDNA-based microarray systems.* Physiol Genomics, 2004. **16**(3): p. 361-70.
25. Lu, J., A. Lal, B. Merriman, S. Nelson, and G. Riggins, *A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips.* Genomics, 2004. **84**(4): p. 631-6.
26. Yauk, C.L. and M.L. Berndt, *Review of the literature examining the correlation among DNA microarray technologies.* Environ Mol Mutagen, 2007. **48**(5): p. 380-94
27. Lee, J.K., K.J. Bussey, F.G. Gwadry, W. Reinhold, G. Riddick, S.L. Pelletier, S. Nishizuka, G. Szakacs, J.P. Annereau, U. Shankavaram, S. Lababidi, L.H. Smith, M.M. Gottesman, and J.N. Weinstein, *Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells.* Genome Biol, 2003. **4**(12): p. R82.
28. Allocco, D.J., I.S. Kohane, and A.J. Butte, *Quantifying the relationship between co-expression, co-regulation and gene function.* BMC Bioinformatics, 2004. **5**(1): p. 18.
29. Kim, S.K., J. Lund, M. Kiraly, K. Duke, M. Jiang, J.M. Stuart, A. Eizinger, B.N. Wylie, and G.S. Davidson, *A gene expression map for Caenorhabditis elegans.* Science, 2001. **293**(5537): p. 2087-92.

30. Jelinsky, S.A., P. Estep, G.M. Church, and L.D. Samson, *Regulatory networks revealed by transcriptional profiling of damaged Saccharomyces cerevisiae cells: Rpn4 links base excision repair with proteasomes.* Mol Cell Biol, 2000. **20**(21): p. 8157-67.
31. Monsieurs, P., G. Thijs, A.A. Fadda, S.C. De Keersmaecker, J. Vanderleyden, B. De Moor, and K. Marchal, *More robust detection of motifs in coexpressed genes by using phylogenetic information.* BMC Bioinformatics, 2006. **7**: p. 160.
32. Quackenbush, J., *Genomics. Microarrays--guilt by association.* Science, 2003. **302**(5643): p. 240-1.
33. Wolfe, C.J., I.S. Kohane, and A.J. Butte, *Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks.* BMC Bioinformatics, 2005. **6**: p. 227.
34. Williams, E.J. and D.J. Bowles, *Coexpression of neighboring genes in the genome of Arabidopsis thaliana.* Genome Res, 2004. **14**(6): p. 1060-7.
35. Mecham, B.H., G.T. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D.Z. Wetmore, T.J. Mariani, I.S. Kohane, and Z. Szallasi, *Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements.* Nucleic Acids Res, 2004. **32**(9): p. e74.
36. Ross, D.T., U. Scherf, M.B. Eisen, C.M. Perou, C. Rees, P. Spellman, V. Iyer, S.S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J.C. Lee, D. Lashkari, D. Shalon, T.G. Myers, J.N. Weinstein, D. Botstein, and P.O. Brown, *Systematic variation in gene expression patterns in human cancer cell lines.* Nat Genet, 2000. **24**(3): p. 227-35.
37. Nacht, M., T. Dracheva, Y. Gao, T. Fujii, Y. Chen, A. Player, V. Akmaev, B. Cook, M. Dufault, M. Zhang, W. Zhang, M. Guo, J. Curran, S. Han, D. Sidransky, K. Buetow, S.L. Madden, and J. Jen, *Molecular characteristics of non-small cell lung cancer.* Proc Natl Acad Sci U S A, 2001. **98**(26): p. 15203-8.
38. Porter, D.A., I.E. Krop, S. Nasser, D. Sgroi, C.M. Kaelin, J.R. Marks, G. Riggins, and K. Polyak, *A SAGE (serial analysis of gene expression) view of breast tumor progression.* Cancer Res, 2001. **61**(15): p. 5697-702.
39. Pruitt, K.D., K.S. Katz, H. Sicotte, and D.R. Maglott, *Introducing RefSeq and LocusLink: curated human genome resources at the NCBI.* Trends in Genetics, 2000. **16**(1): p. 44-47.
40. Mammalian Gene Collection Program Team*, R.L. Strausberg, E.A. Feingold, L.H. Grouse, J.G. Derge, R.D. Klausner, F.S. Collins, L. Wagner, C.M. Shenmen, G.D. Schuler, S.F. Altschul, B. Zeeberg, K.H. Buetow, C.F. Schaefer, N.K. Bhat, R.F. Hopkins, H. Jordan, T. Moore, S.I. Max, J. Wang, F. Hsieh, L. Diatchenko, K. Marusina, A.A. Farmer, G.M. Rubin, L. Hong, M. Stapleton, M.B. Soares, M.F. Bonaldo, T.L. Casavant, T.E. Scheetz, M.J. Brownstein, T.B. Usdin, S. Toshiyuki, P. Carninci, C. Prange, S.S. Raha, N.A. Loquellano, G.J. Peters, R.D. Abramson, S.J. Mullahy, S.A. Bosak, P.J. McEwan, K.J. McKernan, J.A. Malek, P.H. Gunaratne, S. Richards, K.C. Worley, S. Hale, A.M. Garcia, L.J. Gay, S.W. Hulyk, D.K. Villalon, D.M. Muzny, E.J. Sodergren, X. Lu, R.A. Gibbs, J. Fahey, E. Helton, M. Ketteman, A. Madan, S. Rodrigues, A. Sanchez, M. Whiting, A. Madan, A.C. Young, Y. Shevchenko, G.G. Bouffard, R.W. Blakesley, J.W. Touchman, E.D. Green, M.C. Dickson, A.C. Rodriguez, J. Grimwood, J. Schmutz, R.M. Myers, Y.S.N. Butterfield, M.I. Krzywinski, U. Skalska, D.E. Smailus, A. Schnerch, J.E. Schein, S.J.M. Jones, and M.A. Marra, *Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences.* Proc Natl Acad Sci U S A, 2002. **99**(26): p. 16899-903.

41. Robertson, N., M. Oveisi-Fordorei, S.D. Zuyderduyn, R.J. Varhol, C. Fjell, M. Marra, S. Jones, and A. Siddiqui, *DiscoverySpace: an interactive data analysis application.* Genome Biol, 2007. **8**(1): p. R6.
42. Beaudoing, E., S. Freier, J.R. Wyatt, J.M. Claverie, and D. Gautheret, *Patterns of variant polyadenylation signal usage in human genes.* Genome Res, 2000. **10**(7): p. 1001-10.
43. Iseli, C., B.J. Stevenson, S.J. de Souza, H.B. Samaia, A.A. Camargo, K.H. Buetow, R.L. Strausberg, A.J. Simpson, P. Bucher, and C.V. Jongeneel, *Long-range heterogeneity at the 3' ends of human mRNAs.* Genome Res, 2002. **12**(7): p. 1068-74.
44. Unneberg, P., A. Wennborg, and M. Larsson, *Transcript identification by analysis of short sequence tags--influence of tag length, restriction site and transcript database.* Nucleic Acids Res, 2003. **31**(8): p. 2217-26.
45. Davidson, G.S., B.N. Wylie, and K.W. Boyack. *Cluster Stability and the Use of Noise in Interpretation of Clustering*. 2001: IEEE Computer Society. Washington, DC. p. 23.
46. de Hoon, M.J., S. Imoto, J. Nolan, and S. Miyano, *Open source clustering software.* Bioinformatics, 2004. **20**(9): p. 1453-4.
47. Ihaka, R. and R. Gentleman, *R: A language for data analysis and graphics.* Journal of Computational & Graphical Statistics, 1996. **5**(3): p. 299.
48. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
49. Jensen, L.J., J. Lagarde, C. von Mering, and P. Bork, *ArrayProspector: a web resource of functional associations inferred from microarray expression data.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W445-8.
50. Lee, H.K., A.K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, *Coexpression Analysis of Human Genes Across Many Microarray Data Sets.* Genome Res., 2004. **14**(6): p. 1085-94.
51. Bairoch, A., R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.S. Yeh, *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2005. **33**(Database Issue): p. D154-9.
52. Yeung, K., M. Medvedovic, and R. Bumgarner, *From co-expression to co-regulation: how many microarray experiments do we need?* Genome Biol, 2004. **5**(7): p. R48.
53. Carter, S.L., A.C. Eklund, B.H. Mecham, I.S. Kohane, and Z. Szallasi, *Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements.* BMC Bioinformatics, 2005. **6**(1): p. 107.
54. Haverty, P.M., L.L. Hsiao, S.R. Gullans, U. Hansen, and Z. Weng, *Limited agreement among three global gene expression methods highlights the requirement for non-global validation.* Bioinformatics, 2004. **20**(18): p. 3431-41.

3. Implementation and evaluation of Kiwi: A scalable subspace clustering algorithm for the identification of coregulated genes from extremely large gene expression datasets[5,6,7]

3.1. Introduction

Numerous studies have used coexpression of large expression datasets to infer functional associations between genes [1], to identify groups of related genes that are important in specific cancers or represent common tumour progression mechanisms [2], to study evolutionary change [3], for integration with other large-scale datasets [4-6], and for the generation of high-quality biological interaction networks [7-10]. A number of studies have also attempted to use coexpression to identify coregulation with the hypothesis that if two or more genes are expressed at the same time and location and at similar levels then they may be regulated by the same transcription factors and regulatory elements. This approach has shown promise particularly in simpler model organisms such as *A. thaliana* and *S. cerevisiae* [11-14] and many groups are currently working on implementing this idea in mammalian systems. However, traditional clustering methods have not worked particularly well on large datasets for this problem. Most methods assign each gene to only one cluster while in reality many genes likely take part in multiple processes. Also, global coexpression is measured across all conditions, whereas, it is probable that most genes are only tightly coregulated under certain conditions or locations.

In recent years, a new field of clustering analysis termed subspace clustering (or biclustering) has gained increasing popularity in the analysis of gene expression data and other biological data [15-19]. In contrast to traditional clustering methods such as hierarchical clustering, subspace clustering methods do not require expression to be correlated across all conditions for genes to be assigned to the same cluster. This has several advantages for data in which biologically

---

[5] A portion of this chapter has been prepared for publication. Griffith OL, Gao BJ, Bilenky M, Prychyna Y, Ester M and Jones SJM. Implementation and evaluation of Kiwi: A scalable subspace clustering algorithm for the identification of coregulated genes from extremely large gene expression datasets. Manuscript in preparation.

[6] A portion of this chapter has been published. Gao BJ, Griffith OL, Ester M, Jones SJ. 2006. Discovering significant OPSM subspace clusters in massive gene expression data. In Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Philadelphia, PA, USA, August 20-23, 2006). KDD '06. ACM Press, New York, NY, 922-928.

[7] Co-authorship details: I worked with Byron Gao to develop the conceptual requirements of the KiWi subspace clustering algorithm. I was responsible for all biological assessments of the algorithm. Byron Gao wrote and implemented the algorithm itself and performed all computational assessments (not included in this chapter but described in Gao *et al*. (2006)). I was responsible for all analyses, text, figures and tables included in this chapter except where indicated below. The algorithm description (section 3.2.1) was co-written with Byron Gao. Byron Gao also provided Figure 3.1. Mikhail Bilenky developed the promoter similarity score and provided Figures 3.2 and 3.13. Mikhail Bilenky assisted me with the negative control analysis (sections 3.2.8 and 3.3.5; Figures 3.9-3.10) and cisRED analysis (sections 3.2.9 and 3.3.6; Figures 3.11-3.12). Yuliya Prychyna assisted with background research on subspace clustering (Introduction, paragraph 6). Martin Ester and Steven Jones funded and supervised the project.

relevant subsets exist (e.g. different tissue types) or where a few noisy experiments might significantly bias the results of the clustering algorithm. This also allows assignment of genes to multiple clusters for different subsets of experimental conditions.

More recently, the order-preserving sub-matrix (OPSM) has been introduced and demonstrated as a biologically meaningful pattern based subspace cluster model [15, 20]. An OPSM, essentially a subspace cluster, is a subset of rows and columns in a data matrix for which all the rows induce the same linear ordering of columns. In terms of gene expression, an OPSM might represent a group of coregulated genes whose expression levels rise and fall synchronously in response to a series of environmental or cellular stimuli.

A recent report has reviewed several of the existing biclustering methods and compared them to conventional hierarchical clustering methods [17]. They found that in general, biclustering methods outperform global methods such as hierarchical clustering. They also showed that OPSM had the highest proportion of clusters with significant enrichment of one or more Gene Ontology (GO) categories and had good correspondence with known pathways according to their analysis of *A. thaliana* metabolic pathways and *S. cerevisiae* protein-protein interaction networks. However, they state that there are considerable performance differences between the tested methods. Performance is a significant factor as the size of the subspace clustering search space (searching all possible subspaces) is nearly infinite and increases exponentially with the size of the dataset to be analyzed. To illustrate the scale of the problem, consider a dataset of 1,000 genes and 1,000 experiments for which we would like to assess subspace coexpression for all possible gene combinations (of two or more genes) and for all possible experimental subsets (of 10 or more experiments). The number of possible subspaces would be determined as follows:

$$\sum_{g=2}^{1000}\binom{1000}{g}\times\sum_{e=10}^{1000}\binom{1000}{e}=1\times10^{602}.$$

Thus, even in this hypothetical dataset which is considerably smaller than many existing datasets (including some analyzed in this chapter) there are many more possible subspaces than the estimated 1 x 10^{80} atoms in the observable universe [21]. Therefore, a complete and timely solution to large subspace clustering problems is likely unattainable and we will continue to rely on heuristics which attempt partial solutions.

As costs of expression analysis continue to decrease, the numbers and sizes of expression datasets have grown at an ever-increasing rate. The Gene Expression Omnibus (GEO) currently holds more than 170,000 samples for more than 100 different organisms [22] and the Stanford Microarray Database (SMD) contains more than 15,000 public experiments for more than 20 organisms [23]. Furthermore, as array designs continue to improve, it has become possible to include probes for essentially all known genes for many species. The development of exon arrays, alternative splicing arrays, whole-genome tiling arrays, and high-throughput tag-sequencing expression technologies increases the size of the problem even further. Thus, with expression data of potentially tens or hundreds of thousands of both rows and columns we need algorithms that can handle not only 'large datasets' but 'massive datasets'.

In our experience, we have found that none of the existing subspace clustering methods scales well to these larger expression datasets. We have attempted to analyze large datasets with a Gibbs sampling based biclustering algorithm called gene expression mining server (GEMS) [18]; an adaptive quality-based clustering algorithm [16]; an OPSM-based (OP-cluster) algorithm[20]; and PrefixSpan [24], one of the fastest sequential mining algorithms. In all cases, there were either built in limits (e.g. OP-cluster is limited to 100 columns and 5,000 rows) or practical limitations in terms of memory or processor requirements that made it impossible to obtain results for our large datasets (Table 1). In particular, those "twig clusters" defined here as clusters having small size (few genes) and naturally large dimensionality (many experiments) incur explosive computational costs and would be completely pruned off by most existing methods. However, it is of particular interest to biologists to determine small groups of genes that are tightly coregulated under many conditions. Some pathways or processes might require only two genes to act in concert. Thus, there is a clear need for subspace clustering methods that can be run on large datasets and detect these smaller clusters.

To address this challenge we have developed a mining framework that discovers significant OPSM subspace clusters from massive datasets [25]. Here we present an open-source software implementation of this algorithm called KiWi (version 1.0) that is capable of running on a number of different biologically relevant datasets ranging from small to very large in size. We have extensively validated the resulting clusters for these datasets and shown that KiWi correctly assigns redundant probes to the same cluster, groups experiments with common experimental

annotations (such as tissue source), differentiates real promoter sequences from negative control sequences, and groups genes which share common biological processes (GO) and common regulatory sequences as defined by both motif-scanning methods (oPOSSUM) and de novo motif prediction methods (cisRED). As subspace clustering methods continue to gain popularity over simpler global clustering methods, simple and scalable software will be needed to handle the challenge of ever-increasing dataset sizes facing the biologist. To this end, we provide source-code and a working executable for KiWi to the bioinformatics community (http://www.bcgsc.ca/platform/bioinfo/ge/kiwi and http://www.cs.sfu.ca/~bgao/personal/).

3.2. Methods

3.2.1. Algorithm

The KiWi algorithm takes as input a standard gene expression data matrix with genes as rows, experiments as columns, and some measure of expression level for each data point. By sorting the gene (row) vectors and replacing the entries with their corresponding experiment (column) labels, the data matrix can be transformed into a sequence database, and OPSM mining can be reduced to a special case of the sequential pattern mining problem [26]. An OPSM subspace cluster is uniquely specified by this sequential pattern and its supporting sequences. In other words, we are looking for a set of genes (the supporting sequences) which have the same linear order of expression values for some subset of experiments (the pattern). A simple example of an OPSM is shown in Figure 3.1. Instead of finding a complete set of patterns that are beyond some minimum number of genes, KiWi targets the longest patterns for any fixed number of genes (i.e. the subspace cluster with the most experiments showing the same pattern of expression).

KiWi exploits two parameters k and w to provide a biased testing on a bounded number of candidates, substantially reducing the search space and problem scale. In particular, KiWi performs a level-wise search, where shorter patterns gradually grow into longer patterns level by level. Based on the observation that more frequent sub-patterns are likely to grow into more frequent super-patterns, we keep the top k patterns at each level with the greatest number of supporting genes. These patterns are used to generate candidates for the next level. Based on the observation that a long pattern segments its supporting sequences into small sections, in counting the number of supporting sequences, we only consider a region of width w. In other words, only if the new element of a candidate appears in the next w positions of a sequence do we consider the candidate to be supported by the sequence. Other techniques employed by KiWi, such as the

choice of ranking statistics, memory management, pattern extension and redundancy removal, are discussed elsewhere in detail [25].

To our knowledge, KiWi is also the first subspace clustering method to identify anti-correlation as well as correlation within the domain of gene expression analysis. Anti-correlation of expression is interesting because it can also imply common process/pathway membership or negative regulation [27]. Anti-correlated genes might also represent members of opposing pathways (when one is active the other is repressed) [28] or cases where expression of one gene represses the expression of other genes (negative regulators). Anti-correlation in the context of subspace clustering can be captured by the so-called generalized order-preserving sub-matrix (GOPSM) [25], where all the genes in a GOPSM induce the same *or opposite* linear ordering of experiments (see Figure 3.1). KiWi mines GOPSM subspace clusters by searching the sequence database forward and backward simultaneously. KiWi marks any cluster that contains one or more anti-correlated genes. If an 'anti-correlated cluster' contains more than two genes it will by definition contain both positively and negatively correlated pairs of genes.

3.2.2. Datasets

Several datasets and methods were utilized to test for biological coherence of gene clusters predicted by the KiWi clustering algorithm. Expression datasets utilized include: (1) GPL96 - a set of 1,640 Affymetrix (HG-U133A) experiments from the Gene Expression Omnibus (GEO, GPL96) [22] covering a broad range of experimental conditions; (2) expO - a set of 1,026 Affymetrix (HG-U133 Plus 2.0) experiments from 123 different cancer tissue types from the expO (Expression Project for Oncology) project (GEO, GSE2109); (3) Cooper promoters - a high-throughput promoter dataset reported by Cooper *et al.* (2006) consisting of 16 cell lines for which the expression of 632 promoter sequences (plus 98 negative control sequences) were assayed by reporter gene assay [29]. These datasets are summarized in Table 1.

3.2.3. Dataset processing

The Affymetrix HG-U133A (GPL96) probes were normalized and mapped to gene/protein identifiers as described in chapter 2. The expO data were normalized from CEL files using the Bioconductor 'just.gcrma' function in the 'gcrma' library (version 2.4.1). Probes were mapped to Uniprot and Ensembl identifiers using the Bioconductor 'biomart' package [30]. The Cooper promoter data were used as provided. Tab-delimited data matrices for each dataset were loaded

into the KiWi software (v. 1.0) and subspace clusters identified using the parameters outlined in Table 2. Parameters were chosen by experimentation to identify values of k and w that would produce the largest number of clusters and longest patterns but still run to completion in 24 to 48 hours.

3.2.4. Gene Ontology analysis

The first validation method uses the Gene Ontology (GO), a set of structured, controlled vocabularies to identify functional associations between gene products [31]. Current GO annotations and external references file were downloaded from the Gene Ontology Annotation resource at EBI (http://www.ebi.ac.uk/GOA/). Each cluster of protein ids was submitted to the High-Throughput GoMiner command-line interface [32]. Statistically over-represented GO terms were defined using a Fisher's exact test and corrected for multiple testing by false discovery rate detection (100 permutations). All computation was done on a 400+ core (CPUs) OSCAR compute cluster running Red Hat Enterprise Linux 4.

3.2.5. oPOSSUM analysis

The second method uses the oPOSSUM tool to identify statistically over-represented transcription factor binding sites (TFBS) [33]. The oPOSSUM API and MySQL database were downloaded and installed locally (http://www.cisreg.ca/cgi-bin/oPOSSUM/oPOSSUM). Each cluster of genes was submitted to the software and statistically over-represented TFBSs were defined using the Z-score option.

3.2.6. Grouping of probes to common gene identifier

Affymetrix gene expression chips (such as the HG-U133 Plus 2.0 platform, used for expO) contain numerous probe sets derived from the same gene either as redundant probe sets or probe sets for different transcripts of the same gene. It is expected that such probe sets will display correlated expression given that they measure the same or related transcripts. Therefore, we expect that probe sets mapped to a common gene identifier will be grouped together in the same subspace cluster more often than expected by chance. This represents a kind of positive control experiment. The number of probe pairs mapped to the same gene was determined for all clusters. Significance was assessed by random permutation analysis. That is, 10,000 sets of clusters were randomly generated (with the same sizes and dimensions as produced by KiWi). The mean

number of redundant probe pairs for all clusters was then compared between KiWi and the distribution of random results.

3.2.7. Experimental annotation analysis

KiWi subspace clusters consist of a set of 2 or more genes found to have correlated expression patterns (specifically an OPSM) for some subset of the available experiments. Most validation methods look for biologically consistent grouping of genes. To determine if the experiments were also grouped in a meaningful way, an over-representation analysis was applied to experimental annotations. The expO dataset was chosen as this dataset is accompanied by carefully annotated clinical details. For example, each experiment is annotated as one of 123 different tumour tissue types. Clusters with a minimum of two genes and 50 experiments were selected for analysis. Statistically over-represented experimental annotation terms were defined using Fisher Exact statistics and corrected for multiple testing by a Benjamini and Hochberg (BH) correction with the Bioconductor 'multtest' package. Overall significance of the KiWi clusters was determined by comparing to randomly generated clusters using a Kolmogorov-Smirnov test.

3.2.8. Negative control analysis

In the Cooper promoter dataset, expression levels were measured by reporter gene assay for a large number of promoter sequences across a set of 16 different cell lines. They also included a large number of negative control sequences (random DNA sequence). Unlike the real promoter sequences, we do not expect these sequences to drive gene expression in any meaningful way. They may produce some low level 'noisy' expression values but nothing coordinated. Any clusters formed from such data most likely represent random patterns as opposed to meaningful patterns. Therefore, we can hypothesize that if Kiwi is detecting 'true' or 'real' clusters they should be biased towards positive sequences and against negative sequences. To this end, we ran KiWi on expression data for all sequences (positives and negatives) and determined the fractions of KiWi clusters 'contaminated' by negative control sequences. Significance was determined using a random permutation approach as described above. We also clustered positive and negative sequences separately and compared the numbers, sizes (number of genes) and pattern lengths (number of experiments) of clusters for each dataset.

3.2.9. cisRED analysis

For all Cooper promoter sequences (excluding negative control sequences) the cisRED pipeline was used to predict putative regulatory motifs as described previously [34]. Briefly, cisRED uses multiple discovery methods applied to sequence sets that include up to 42 orthologous sequence regions from vertebrates (mostly mammals). Motif significance is estimated by applying discovery and post-processing methods to randomized sequence sets that are adaptively derived from target sequence sets. Motifs are then annotated based on their similarity to known transcription factor binding site (TFBS) models (using TRANSFAC 9.3). Motifs with p-values below a threshold (discovery p-value < 0.001 and annotation p-value < 0.0005) are retained, groups of similar motifs identified and co-occurring motif patterns defined. Any set of two or more genes with a co-occurring motif pattern is hypothesized to be co-regulated by one or more transcription factors. We can further hypothesize that these putatively co-regulated gene groups are more likely to belong to a cluster of coexpressed genes (as defined by KiWi) than randomly formed groups of genes. To test this, a promoter similarity score was defined. For every KiWi cluster, for each pair of genes in the cluster, we calculate the number of cisRED motifs annotated with the same TFBS model (counting repeated annotations only once; Figure 3.2). More formally, we define $L_i = \{l_{i,}^1, l_i^2, ..., l_i^n\}$ to be a set of annotation labels for n conserved motifs predicted in the promoter of a gene "i". Then we define a promoter similarity score (S) as the average number of common annotation labels for every pair (i,j) of genes from a KiWi cluster of size N as follows:

$$S = 2\frac{\sum_{i,j} | L_i \cap L_j |}{N(N-1)}.$$

3.2.10. Supplementary materials

All other data files necessary to reproduce the analysis are available on the supplementary website at: http://www.bcgsc.ca/platform/bioinfo/ge/kiwi.

3.3. Results

3.3.1. KiWi subspace clustering results

For all three datasets analyzed, KiWi was able to run to completion (after parameter optimization) and produce a large number of clusters. The results are summarized in Table 3.3 and density distributions for cluster size (number of genes) versus pattern length (number of

experiments) plotted in Figure 3.3. A large number of clusters (13,412 to 212,532) were identified for the three datasets, with a range of sizes and pattern lengths. In general, KiWi appears well suited to identifying smaller clusters with long patterns. For the GPL96, expO, and Cooper datasets the average cluster size was 5.11, 3.89, and 6.79 and the average pattern length was 24.04, 42.48, and 6.85 respectively.

3.3.2. GO and oPOSSUM analysis

GO and oPOSSUM analysis results are shown for the GPL96 dataset (Figures 3.4 and 3.5). For the validation, a set of representative clusters were chosen with minimum size 5 and minimum dimensions 15. A total of 634 clusters met these criteria, with 22 clusters containing anti-correlated genes. The GO analysis shows that clusters identified by KiWi are significantly more likely to share a common biological process than random expectation (Figure 3.4). For example, if we consider a p-value threshold of 0.01, more than 10% of clusters have at least one significant GO term compared to the random expectation of close to zero. Similarly, the TFBS analysis shows that clusters identified by the KiWi algorithm are significantly more likely than random expectation to share sequences bound by the same transcription factor (Figure 3.5). For example, if we consider a Z-score of 30, more than 10% of clusters have at least one TFBS over-represented in the regulatory regions for these genes. Random expectation for this same Z-score threshold is close to zero. To summarize, the KiWi algorithm successfully identifies groups of genes that belong to a common function or process and/or share common transcription factor binding sites.

3.3.3. Grouping of probes to common gene identifier

Figure 3.6 shows the results of the 'probe to common gene' analysis. Using the expO dataset (because it is based on the more current Affymetrix platform) we found that of the 23,705 clusters identified by KiWi, 1,880 (7.93%) contained at least one pair of redundant probes (i.e. different probes corresponding to the same gene) and on average, KiWi clusters contained 0.177 redundant probe pairs per cluster. This was significantly more than the 24.5 (0.10%) clusters with at least one redundant pair ($p<0.0001$, 10,000 permutations) and the 0.002 average number of redundant probes per cluster ($p<0.0001$, 10,000 permutations) identified in our random simulations.

3.3.4. Experimental annotation analysis

Figure 3.7 shows the results of the 'experimental annotation analysis'. This used the expO data because unlike most gene expression datasets, the expO data are accompanied by careful and comprehensive experimental annotations. The graph shows a significant tendency by KiWi to group experimental dimensions with common experimental annotation terms such as tissue source, histology, gender, ethnicity, smoking, or alcohol consumption status ($p=0.009$). When only tissue source terms were considered, a nearly identical graph was observed with a similar level of significance ($p=0.005$, Figure 3.8).

3.3.5. Negative control analysis

Figure 3.9 shows the average number of negative control sequences (for the Cooper promoter dataset) in KiWi clusters for different pattern lengths (number of experiments) and cluster size (number of genes/promoters). The random expectation is that negative sequences will be included at a constant rate based on the proportion of total genes/promoters that are negatives. This is what we observe with the randomly generated clusters having a very constant mean fraction of negative controls for all cluster sizes and pattern lengths. Overall, the fraction of negative control sequences included in KiWi clusters was 0.129. This was significantly lower than the mean fraction of 0.134 observed for random simulations ($p<0.001$, 1,000 permutations). The more genes that form a cluster (share a KiWi pattern), the less likely that cluster is to include negative control sequences. Similarly, the longer the pattern (more experimental dimensions) a cluster has, the less likely that cluster is to include a negative control sequence. When negative control sequences were excluded and clustered separately from real sequences we found that significantly more clusters (with greater cluster size and patterns lengths) were produced for the real data than the negative control data (Figure 3.10).

3.3.6. cisRED analysis

Figure 3.11 shows that KiWi clusters for the Cooper promoter dataset are more likely to have promoters that contain similar conserved motifs than randomly grouped genes. The overall mean promoter similarity score for KiWi clusters of 0.339 was significantly greater than the score of 0.296 for random (Wilcoxon test, $p < 2.2e\text{-}16$). As the cluster size increases (more genes) the separation from random increases. This is also true for increasing number of experiments (Figure 3.12).

3.4. Discussion

We have extensively assessed and validated the performance of the first subspace clustering implementation (for gene expression analysis) that is scalable to very large datasets (10,000s of genes and 1,000s of experiments) and able to identify smaller (twig) clusters. An advantage of KiWi is that subspace clusters can be identified for a dataset of virtually any size. By experimenting with the settings for k and w, the user can balance the number and quality of clusters identified against desired runtime. We chose settings that would produce the best results (in terms of numbers of clusters and maximum pattern length) but still run to completion in ~24-48 hours on an ordinary PC. Typically subspace clustering methods are evaluated on significantly smaller datasets and with only preliminary biological assessments such as GO analysis [17]. Here, we go significantly further and demonstrate KiWi's ability to identify biologically interesting clusters (at both the gene level and experiment level) from expression datasets that were, in our experience, inaccessible to available subspace clustering implementations in the gene expression analysis field.

An initial matter for discussion is the idea that KiWi is able to identify smaller clusters (the so-called twig clusters) rather than being limited to clusters with very large numbers of genes. If we plot a density graph of the number of genes versus number of experiments for all clusters we see that there is a bias towards smaller clusters. For example, in the expO data 80% of clusters had 5 genes or less and 97% had 10 genes or less (Figure 3.3B) although some larger clusters were found as well (up to 162 genes in the GPL96 dataset, 23 in the expO dataset, and 69 in the Cooper dataset). These clusters, while small in gene number, are in many cases coexpressed across a large number of experiments. For clusters with 5 genes or less, the number of experiments over which coexpression was observed ranged from 10 to 120 with a mean of 42.48. Similar trends were seen for the GPL96 and Cooper data (Table 3; Figure 3.3). We believe this is a novel contribution to the subspace clustering field. Previous studies have tended to focus on the larger clusters by design or necessity. While these large clusters have been shown to be of interest or value, there is no reason to expect that all or even most biological processes or disease mechanisms would involve tight co-regulation of large groups of genes. In fact, we expect that many important processes will involve relatively small numbers of genes. Indeed, the biological assessments discussed below showed that many of these small clusters are of biological interest.

Gene Ontology (GO) and oPOSSUM performance were evaluated using a large Affymetrix dataset (referred to here as GPL96). We showed that KiWi can group genes with over-representation of GO biological processes and oPOSSUM transcription factor binding sites. In a previous study (see Chapter 2) [35] we found that for the GPL96 data, approximately 33% of coexpressed gene pairs with a global Pearson correlation (r) of at least 0.80 shared a common GO biological process term. This was the minimum r value for which we saw clear separation from random performance. Other studies have recommended r=0.84 as a good cutoff for reliable global coexpression [36]. However, even with the less stringent cutoff of r=0.8, only 9701 gene pairs actually attained this level of global coexpression. We hypothesized that subspace clustering might provide a useful alternative or complementary method for identifying biologically relevant coexpression relationships. With KiWi we were able for the first time to analyze our large GPL96 dataset (updated with a further 973 experiments since the previous study) for subspace coexpression. The entire set of 13,412 KiWi clusters (representing 393,352 coexpressed gene pairs) had similar levels of GO performance with 23% of clusters having at least one significant GO term (FDR<0.1) and good separation from random expectation. This is perhaps expected given that the vast majority of KiWi clusters have very high correlation of expression between genes across their subset of experiments. For example, in the GPL96 dataset, 90% of KiWi clusters had an average r > 0.95 (Figure 3.13) whereas only 44 of the gene pairs identified by global methods had r values this high. This demonstrates that whereas few gene pairs or clusters are highly correlated across all conditions/experiments, most genes are highly correlated with one or more other genes across some subset of the conditions.

Using another large Affymetrix dataset (expO) we show that KiWi is also very good at grouping probes for the same gene. In fact, a significant proportion of the total clusters contain 'redundant' probes. In itself, this is not a surprising result, but is an important positive control that shows KiWi correctly identifies logically related genes. However, this also argues for removing or averaging of redundant probes before clustering to avoid wasted computation time when gene clusters (as opposed to probe clusters) are the desired end-product.

The expO dataset was also useful for its evaluation of experiment-level clustering by KiWi. Part of the promise of biclustering methods is that they will identify not only coexpressed genes but also the subset of experimental tissues or conditions under which genes are coexpressed. Such biclusters could be of particular value for identifying tissue- or stage-specific coregulation. But,

they also allow identification of coregulation for previously unconsidered sample groups. For example, we were able to identify gene clusters specific to gender, smoking and alcohol consumption status. The Expression Project for Oncology and International Genomics Consortium should be applauded for not only making their raw expression data (CEL files) available in GEO but also for providing these detailed and standardized clinical annotations. Almost none of the other datasets in GEO have done so. Such datasets make possible, for the first time, true two-dimensional evaluation of biclustering methods on clinically relevant expression data.

In the Cooper promoter dataset, expression levels were measured by reporter gene assay for a large number of promoter sequences across a set of 16 different cell lines. Thus, the data comes in the normal format of a gene/promoter versus experiment/condition matrix but expression is measured by reporter gene activity (luciferase levels) instead of hybridization intensity (as on a microarray). The Cooper promoter dataset is a relatively small dataset and was chosen for reasons other than its size. However, in subspace clustering problems, even a seemingly small dataset contains a very large set of possible subspaces. For example, for the Cooper dataset with the parameters we chose, there are 3.3×10^{224} possible subspace clusters. What makes the Cooper dataset particularly useful is the presence of a large number of negative control sequences (random DNA sequence). We hypothesized that KiWi would be biased against inclusion of these negative control sequences and this is indeed what we observed. Also, both the pattern length and number of genes seem to be strong predictors of how reliable a pattern is. We can in principle use this information to define rules for the minimum length and cluster size in combination needed for a 'reliable' pattern. Visually, a cluster with pattern length of 10+ looks reliable with 3 or more genes, a cluster with pattern length of 9 looks reliable with 4 or more genes, and so on.

Comparing the Cooper promoter KiWi clusters to a cisRED analysis of the Cooper promoter sequences also showed a tendency for coexpressed genes to share common regulatory motifs. This result suggests that KiWi coexpression analysis may be useful in filtering de novo motif predictions and/or selecting coexpressed genes as input for motif discovery. As in the negative control analysis, we found that both the cluster size and number of dimensions were predictors of promoter similarity. These findings confirm the intuitive idea that OPSM patterns are more likely to be real if they are shared by more genes and/or across more experiments. This implies that

smaller (twig) clusters will need longer patterns (more experiments) in order to have the same level of confidence as larger clusters. A useful future development of KiWi would be the development of a score or p-value by which clusters could be automatically ranked that takes both the numbers of genes and experiments into consideration.

3.5. Conclusions

By design, KiWi is capable of identifying both negative and positive correlations of expression, twig clusters (as small as two genes), and genes that appear in multiple clusters. We have demonstrated that these clusters correctly group related probe sequences, avoid 'contamination' by negative controls and tend to share common biological processes (GO) and common regulatory sequences as defined by both motif-scanning methods (oPOSSUM) and de novo motif prediction methods (cisRED). Finally, over-representation of experimental annotation terms gives hope that tissue- or condition-specific clusters can be defined. These features suggest that KiWi should be useful for a wide range of biological applications and may be of particular use in the identification of novel groups of coregulated genes. To facilitate these applications, we provide all datasets, source code, a software tutorial and a working executable (for Windows® operating systems) to the bioinformatics research community (http://www.bcgsc.ca/platform/bioinfo/ge/kiwi and http://www.cs.sfu.ca/~bgao/personal/).

Table 3.1. Datasets used in KiWi assessment

Dataset	# of rows	# of columns
GPL96	12,332	1,640
expO	20,113	1,026
Cooper promoters	730	16

Table 3.2. Parameters used in KiWi assessment

Dataset	k	w	Min. genes	Min. exps
GPL96	30,000	45	2	10
expO	100,000	18	2	10
Cooper promoters	100,000	16	2	6

Table 3.3. KiWi results

Note that the minimum values seen for the ranges of numbers of genes and experiments were set in KiWi as parameters (see Table 2).

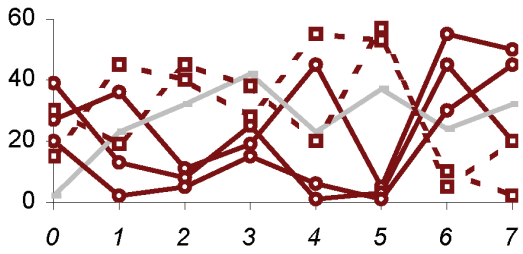
Dataset	# clusters found	Mean genes/cluster (range)	Mean exps/cluster (range)
GPL96	13,412	5.11 (2 to 162)	24.04 (11 to 108)
expO	23,555	3.89 (2 to 23)	42.48 (10 to 120)
Cooper promoters	212,532	6.79 (2 to 59)	6.85 (6 to 14)

Figure 3.1. The order preserving submatrix (OPSM) and generalized OPSM (GOPSM)

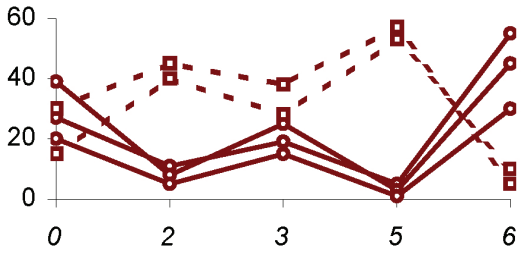
A simple data matrix in (a), representing genes (rows) I through VI and experiments (columns) 0 through 7 each with some expression value, exhibits no obvious pattern when plotted in (b). However, when only a subset of experiments (0, 2, 3, 5, and 6) is considered in (c), there are two order preserving submatrices (OPSMs) or "clusters" clearly present (one for the shaded values and one for the underlined values in the data matrix). A single gene (grey solid line) belongs to neither OPSM. One OPSM (dashed red lines) is perfectly anti-correlated with the other (solid red lines). Therefore, both OPSMs could be combined into a single generalized OPSM (GOPSM) if we group coexpression and anti-coexpression. Finally, (d) shows a permutation of columns of the GOPSM, under which the row sequences are in either strictly ascending or descending order. This more clearly illustrates the linear order of expression values that defines the GOPSM and each OPSM.

	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*
I	27	36	11	19	45	5	55	50
II	39	13	8	25	1	3	45	20
III	2	23	32	42	23	37	24	32
IV	30	19	45	38	20	57	5	20
V	15	45	40	28	55	53	10	2
VI	20	2	5	15	6	1	30	45

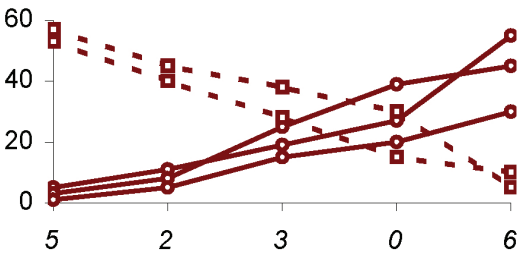
(a) Data matrix



(b) Data matrix plotted



(c) GOPSM consisting of two OPSMs



(d) GOPSM rearranged

Figure 3.2. Diagrammatic explanation of "promoter similarity score"

For any pair of genes, the upstream region is compared for overlap of annotated TFBS motifs (Sp1, AP-2, etc). These are represented as coloured shapes in the diagram below. Each common motif is counted once. The promoters above share two motifs (red square and blue circle). Therefore, the score for this gene pair is 2. Then, the overall promoter similarity score (S) for a cluster of genes is calculated as the sum of the pairwise scores divided by the number of pairs.

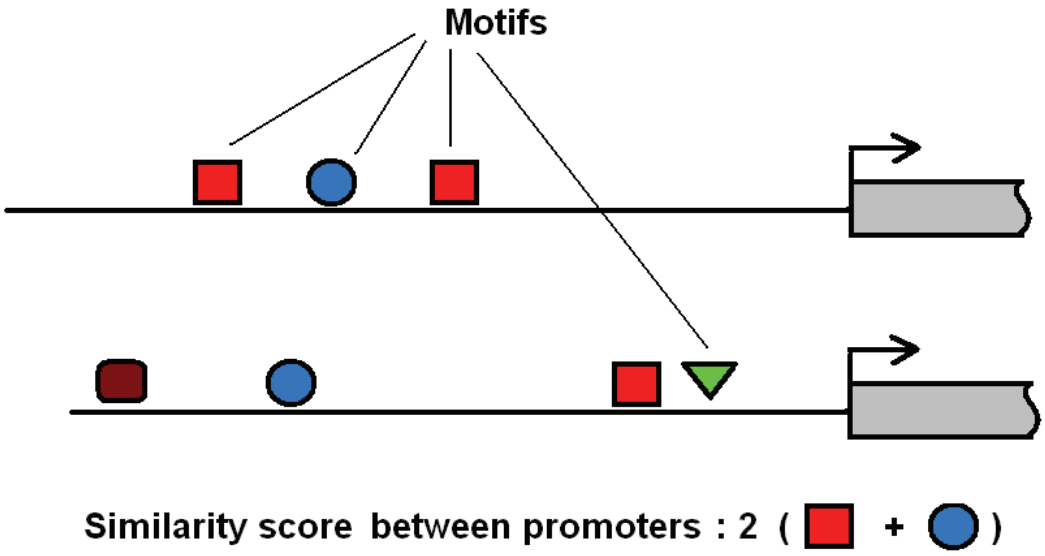


Figure 3.3. KiWi results for the GPL96 dataset

The density plot shows the numbers of clusters for different cluster sizes (number of genes) and pattern lengths (number of experiments) for the GPL96 (A), expO (B) and Cooper promoter (C) datasets. Many clusters were identified for the three datasets, with a range of sizes and pattern lengths. In general, KiWi appears well suited to identify smaller clusters with long patterns. For example, in the expO data (panel B) 80% of clusters had 5 genes or fewer and 97% had 10 genes or fewer although some larger clusters were found as well (up to 23 genes in the expO dataset). These clusters, while small in gene number, are in many cases coexpressed across a large number of experiments. For clusters with 5 genes or fewer, the number of experiments over which coexpression was observed ranged from 10 to 120 with a mean of 42.48. Similar trends were seen for the GPL96 and Cooper data (panels A and C). The density plot was produced using the Bioconductor 'hexbin' library (version 2.3.0).

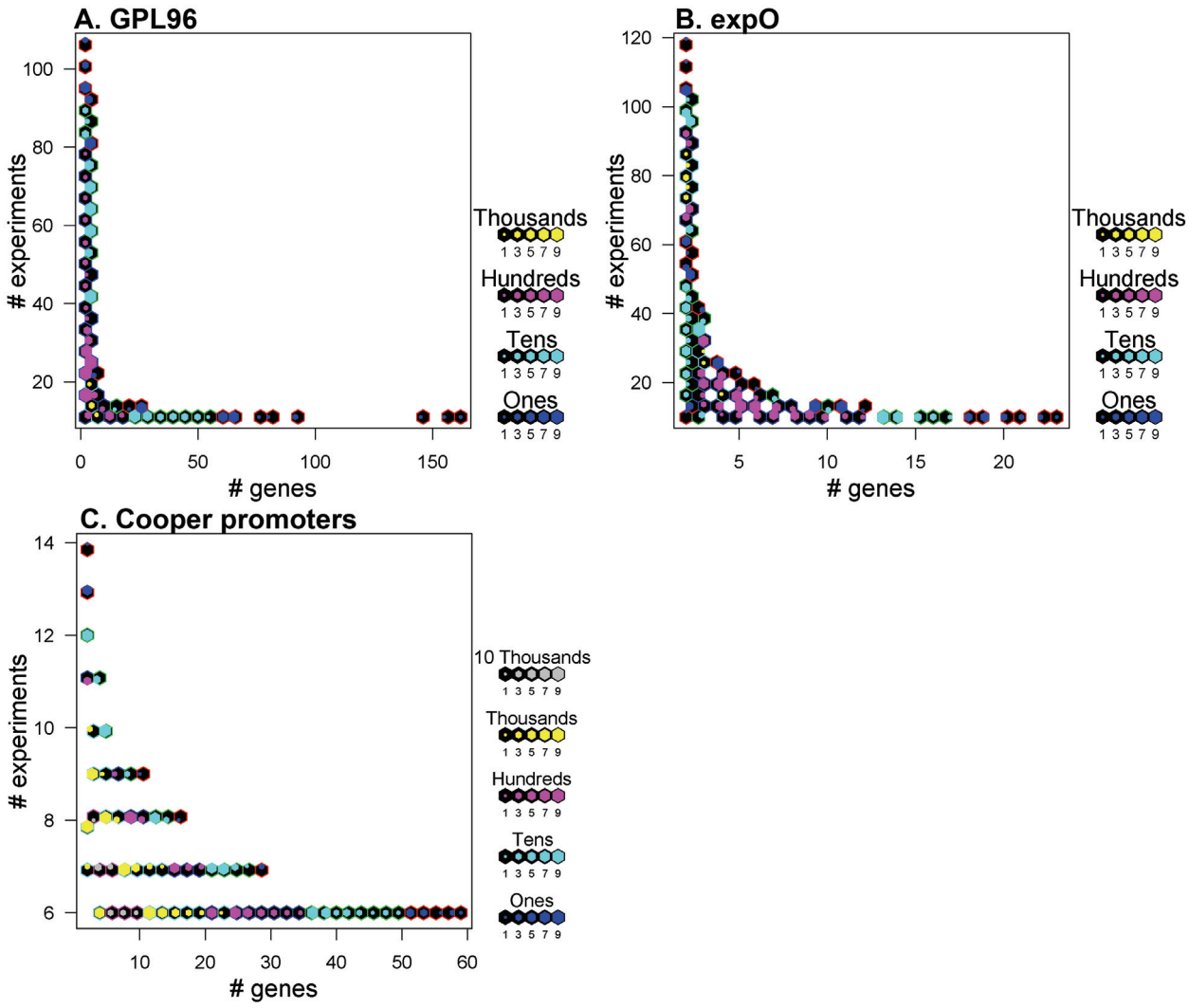


Figure 3.4. Gene Ontology analysis

The fraction of clusters with at least one significantly over-represented GO biological process term at each level of significance is shown. The GO analysis shows that clusters identified by KiWi are significantly more likely to share a common biological process than random expectation. For example, if we consider a p-value threshold of 0.01, more than 10% of clusters have at least one significant GO term compared to the random expectation of close to zero. GO tests were performed by High-throughput GOminer. P-values were FDR corrected (100 permutations) and displayed on a log-scale.

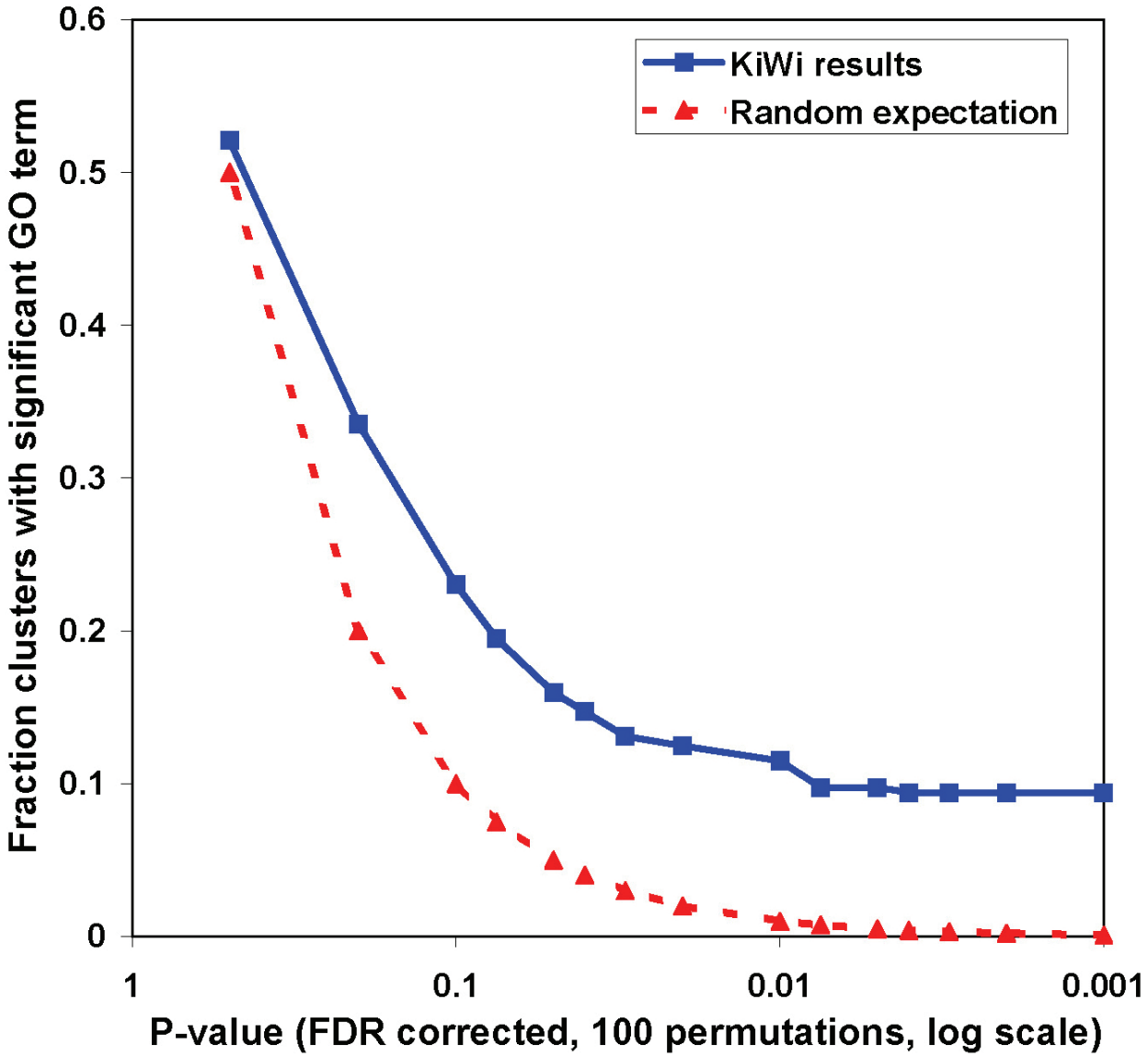


Figure 3.5. oPOSSUM TFBS analysis

The fraction of clusters with at least one significantly over-represented transcription factor binding site (TFBS) at each level of significance is shown. The TFBS analysis shows that clusters identified by the KiWi algorithm are significantly more likely than random expectation to share sequences bound by the same transcription factor. For example, if we consider a Z-score of 30, more than 10% of KiWi clusters have at least one TFBS over-represented in the regulatory regions for these genes. Random expectation for this same Z-score threshold is close to zero. TFBS tests were performed by the oPOSSUM tool using the z-score measure of significance.

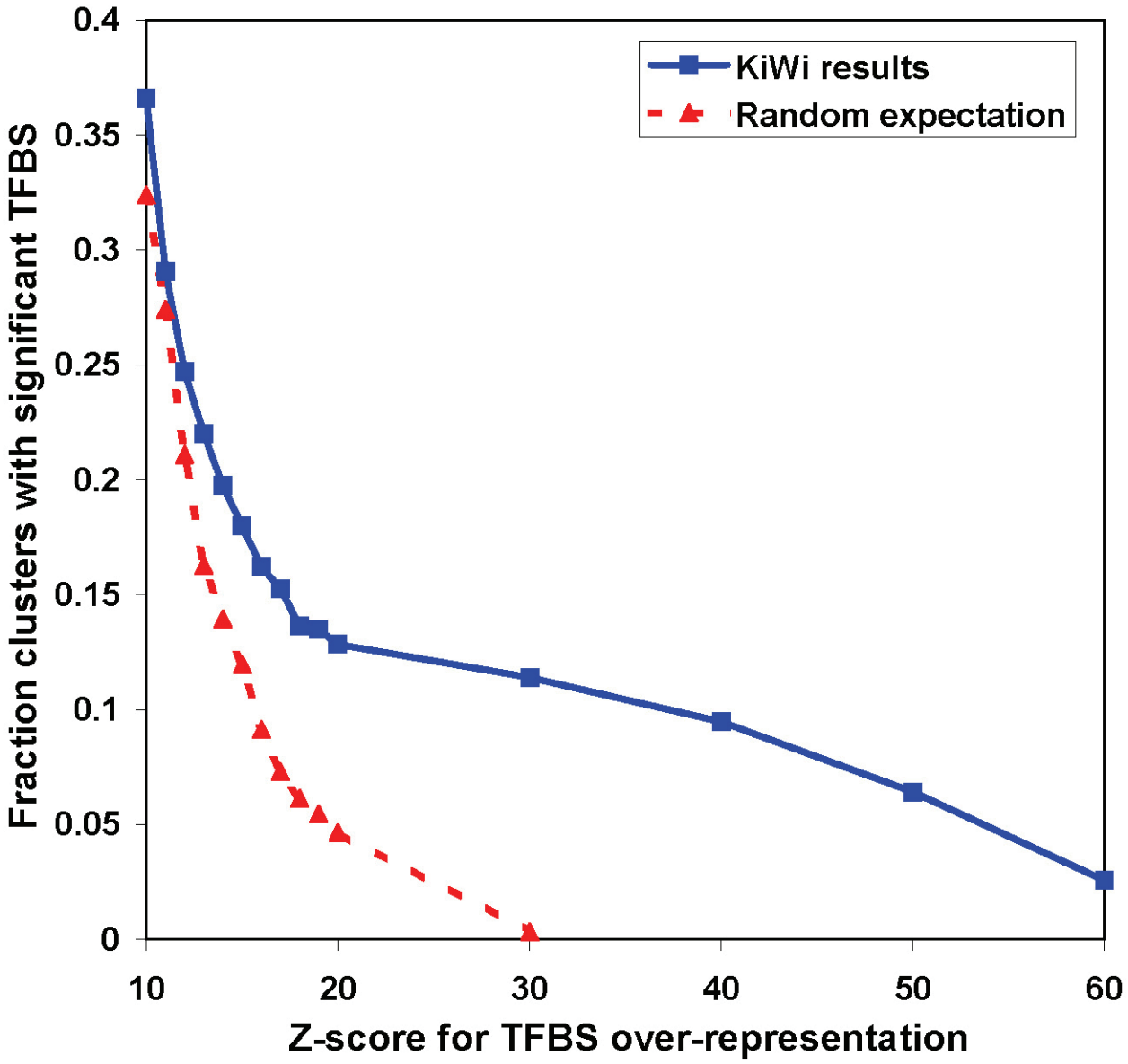


Figure 3.6. Probe to common gene analysis for expO dataset

The mean number of probe pairs in a cluster that are mapped to the same gene (redundant probes) is shown for each cluster size (number of genes per cluster) for the expO dataset. We found that KiWi clusters contained significantly more redundant probe pairs per cluster than our random simulations ($p<0.0001$, 10,000 permutations). Error bars indicate 95% confidence limits.

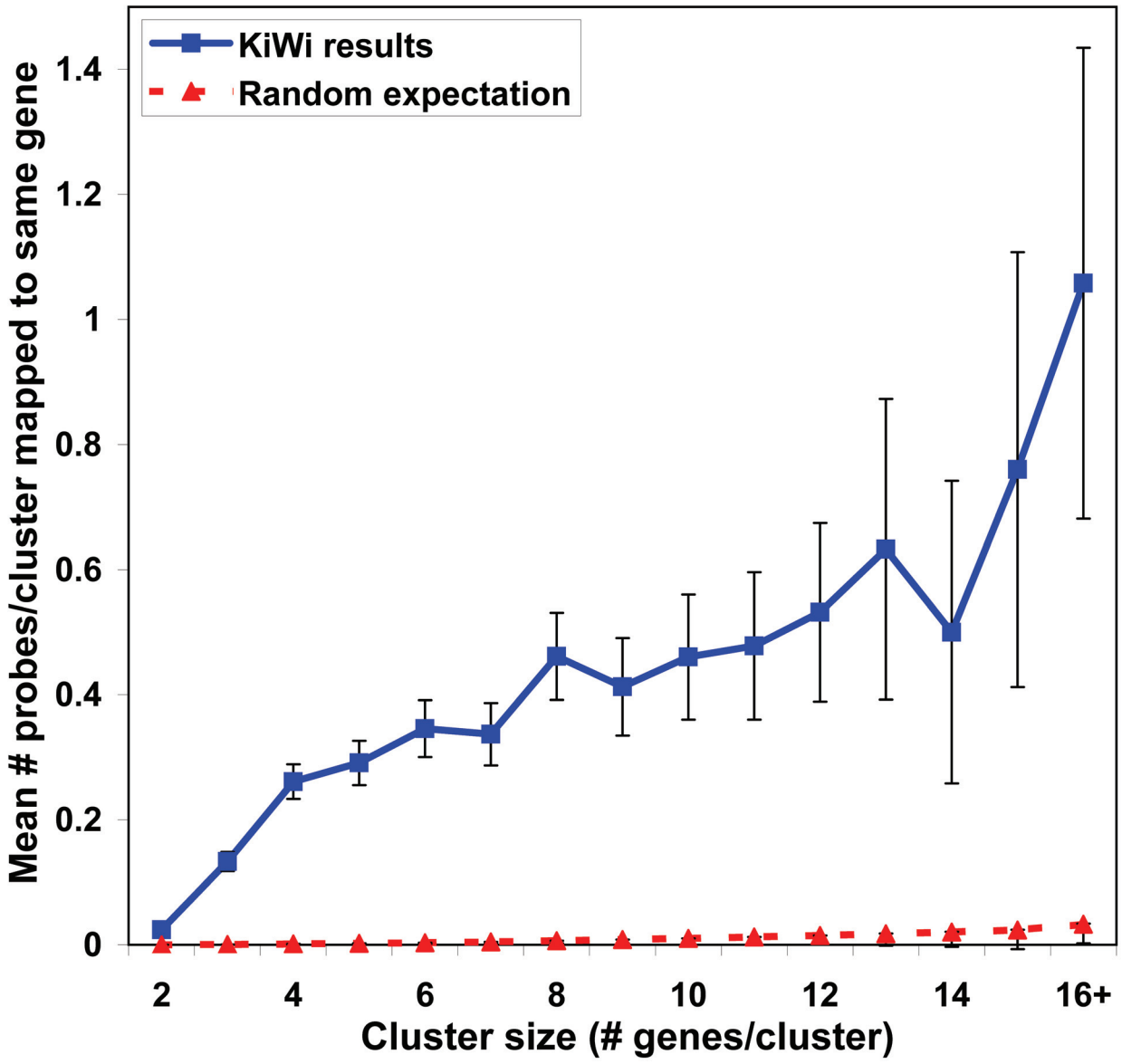


Figure 3.7. Experimental annotation analysis for expO dataset (all annotations)

The fraction of clusters with at least one significantly over-represented experimental annotation term at each level of significance is shown. Kiwi showed a significant tendency to group experiments with common experimental annotation terms such as tissue source, histology, gender, ethnicity, smoking, or alcohol consumption status (p=0.009). Significance for each individual test (i.e., cluster vs annotation term) was determined by Fisher Exact test. P-values were corrected by the Benjamini and Hochberg method and are displayed on a reverse log scale. Significance between Kiwi and random was determined by Kolmogorov-Smirnov test.

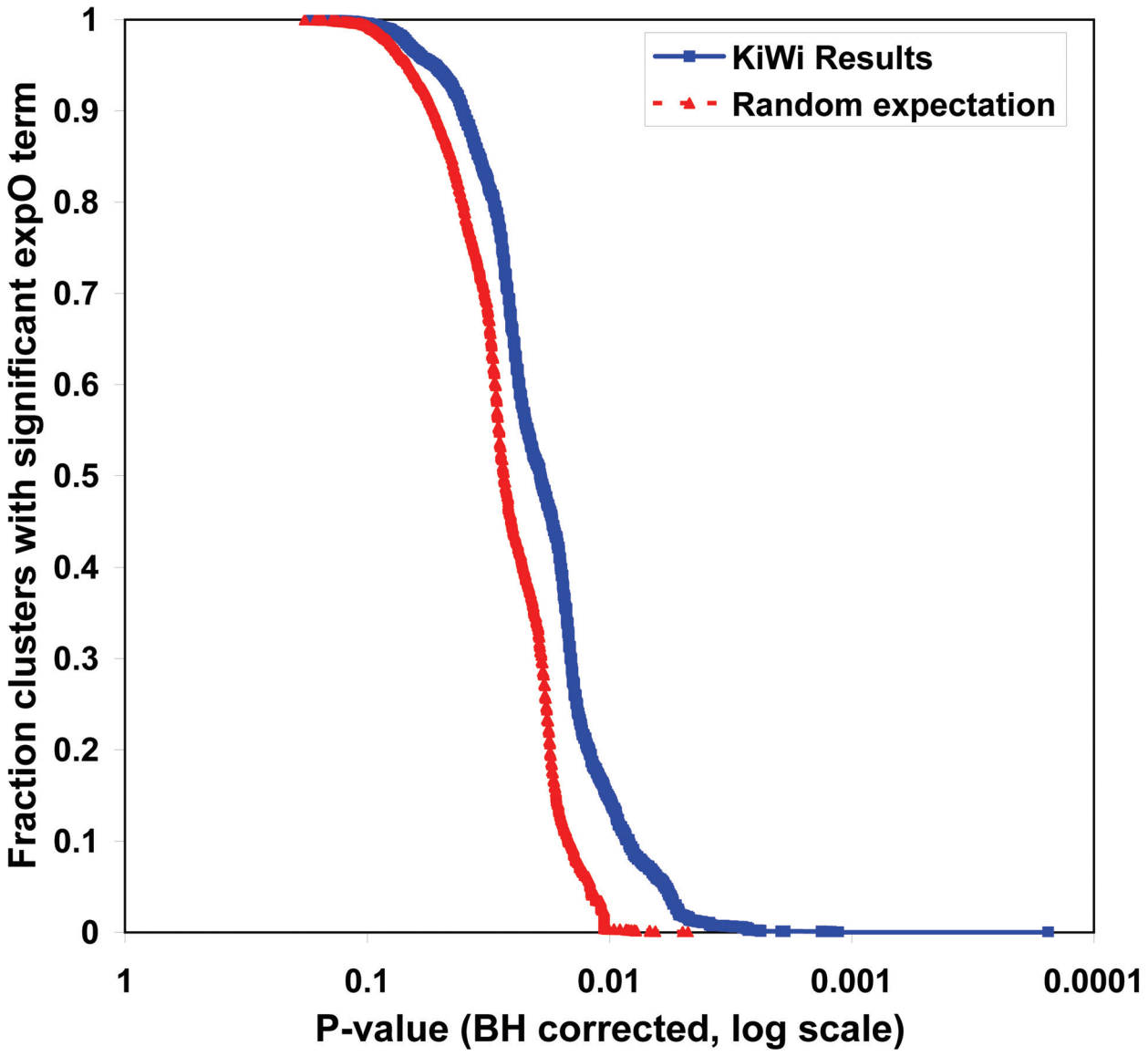


Figure 3.8. Experimental annotation analysis for expO dataset (tissue source only)

The fraction of clusters with at least one significantly over-represented tissue source term at each level of significance is shown. Kiwi showed a significant tendency to group experiments with a common tissue source (p=0.005). Significance for each individual test (i.e., cluster vs annotation term) was determined by Fisher Exact test. P-values were corrected by the Benjamini and Hochberg method. Significance between Kiwi and random was determined by Kolmogorov-Smirnov test.

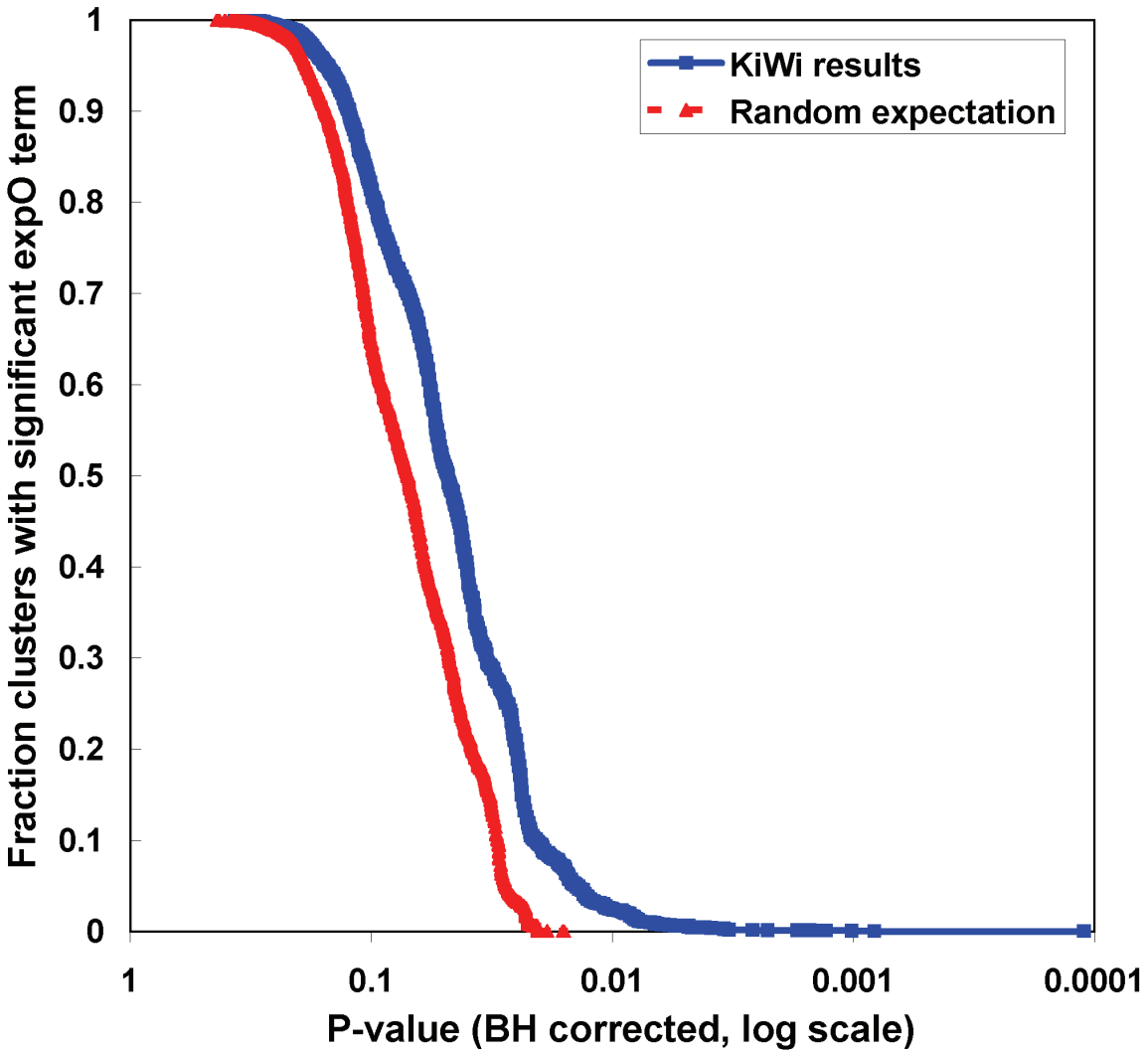


Figure 3.9. Negative control analysis for Cooper promoter dataset

The mean fraction of negative control sequences included in each cluster for each cluster size is shown. Results are broken down by pattern length except for the random results for which all cluster sizes and pattern lengths showed constant contamination by negative controls. The overall fraction of negative control sequences included in KiWi clusters was significantly lower than the mean fraction observed for random simulations ($p<0.001$, 1,000 permutations). In general, the more genes that form a cluster (share a KiWi pattern), the less likely that cluster is to include negative control sequences. Similarly, the longer the pattern (more experimental dimensions) a cluster has, the less likely that cluster is to include a negative control sequence. Error bars indicate 95% confidence limits.

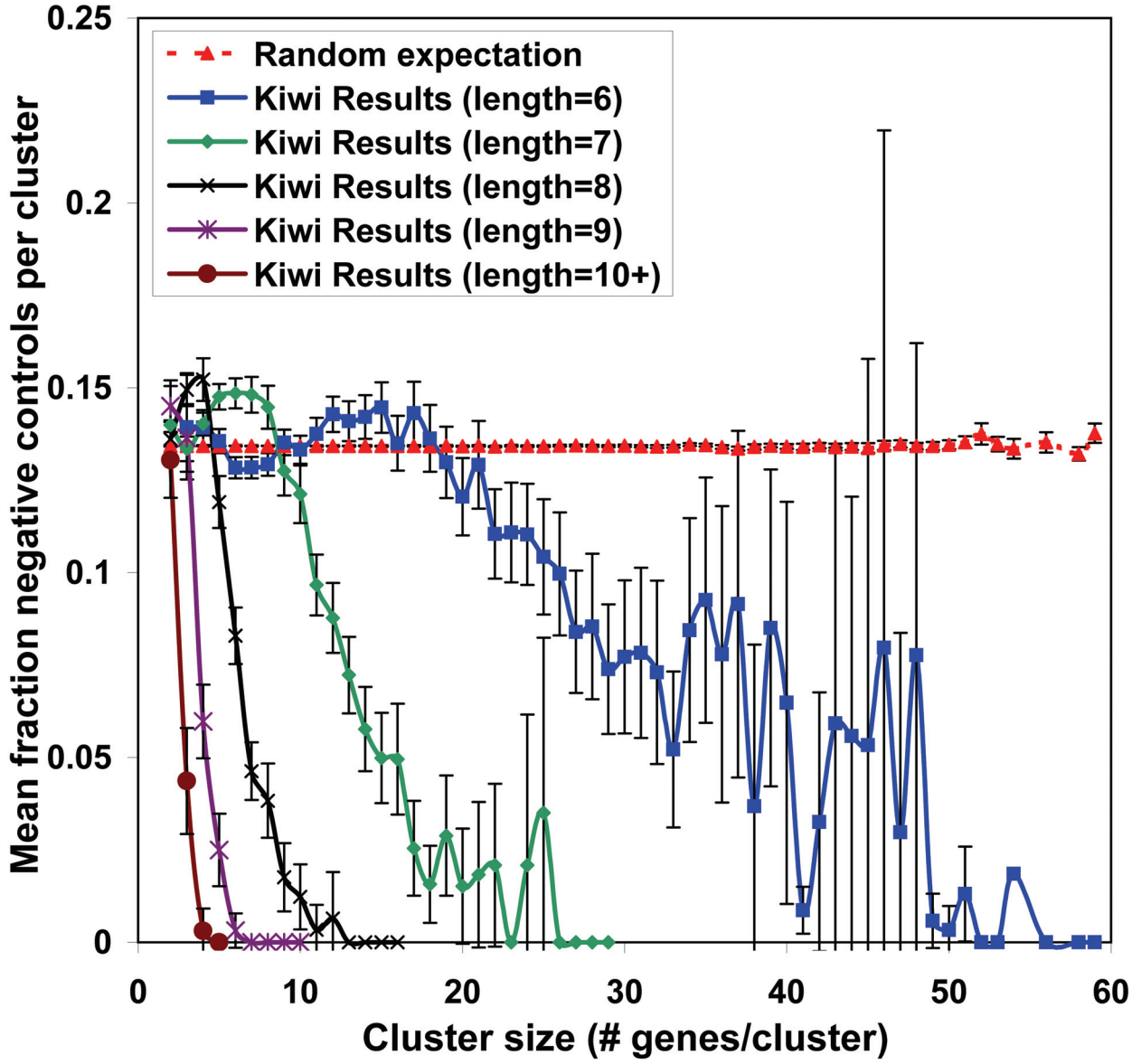


Figure 3.10. KiWi results for Cooper promoter dataset with negative control sequences and real promoter sequences clustered separately

When negative control sequences (Negatives) were excluded and clustered separately from real sequences (Positives) we found that significantly more clusters (with greater cluster size and patterns lengths) were produced for the real data than the negative control data. The density plot was produced using the Bioconductor 'hexbin' library (version 2.3.0).

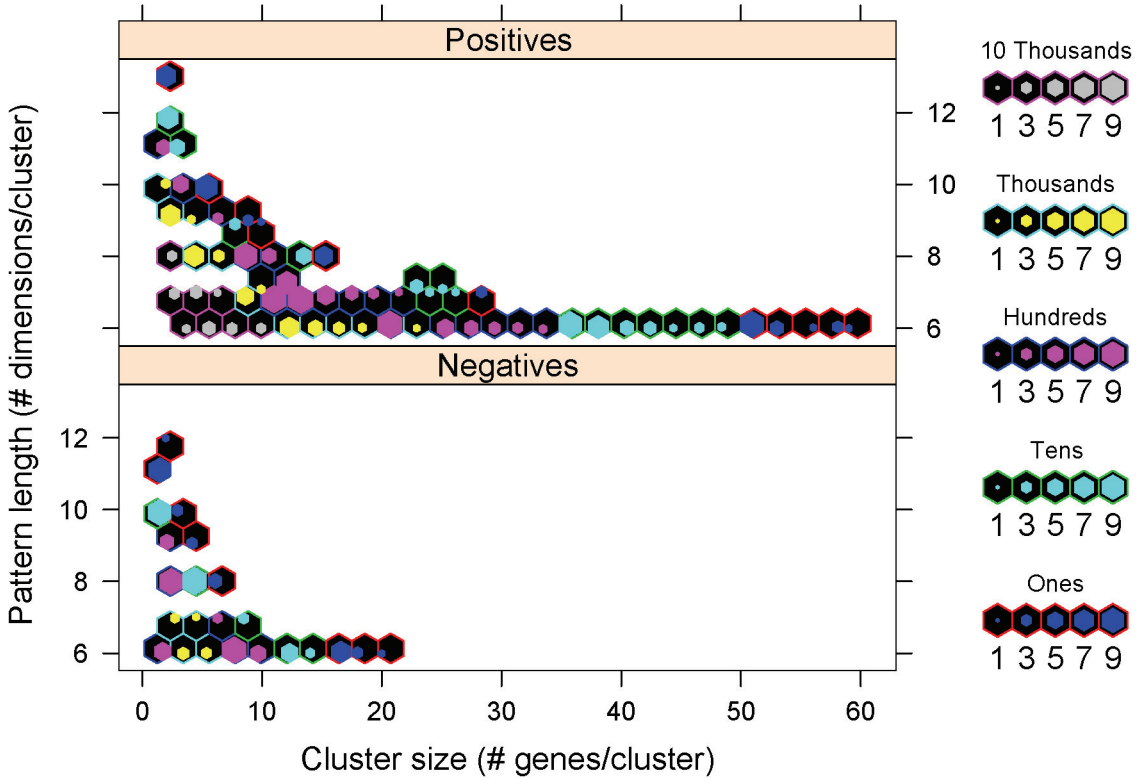


Figure 3.11. cisRED analysis for Cooper promoter dataset

The mean promoter similarity score (see methods) for each cluster size is shown. KiWi clusters for the Cooper promoter dataset are more likely to have promoters that contain similar conserved motifs than randomly grouped genes. The overall mean promoter similarity score for KiWi clusters was significantly greater than the score for random (Wilcoxon test, $p < 2.2e-16$). As the cluster size increases (more genes) the separation from random increases. Random expectation is based on 2000 randomly generated clusters for each cluster size. Error bars indicate 95% confidence limits.

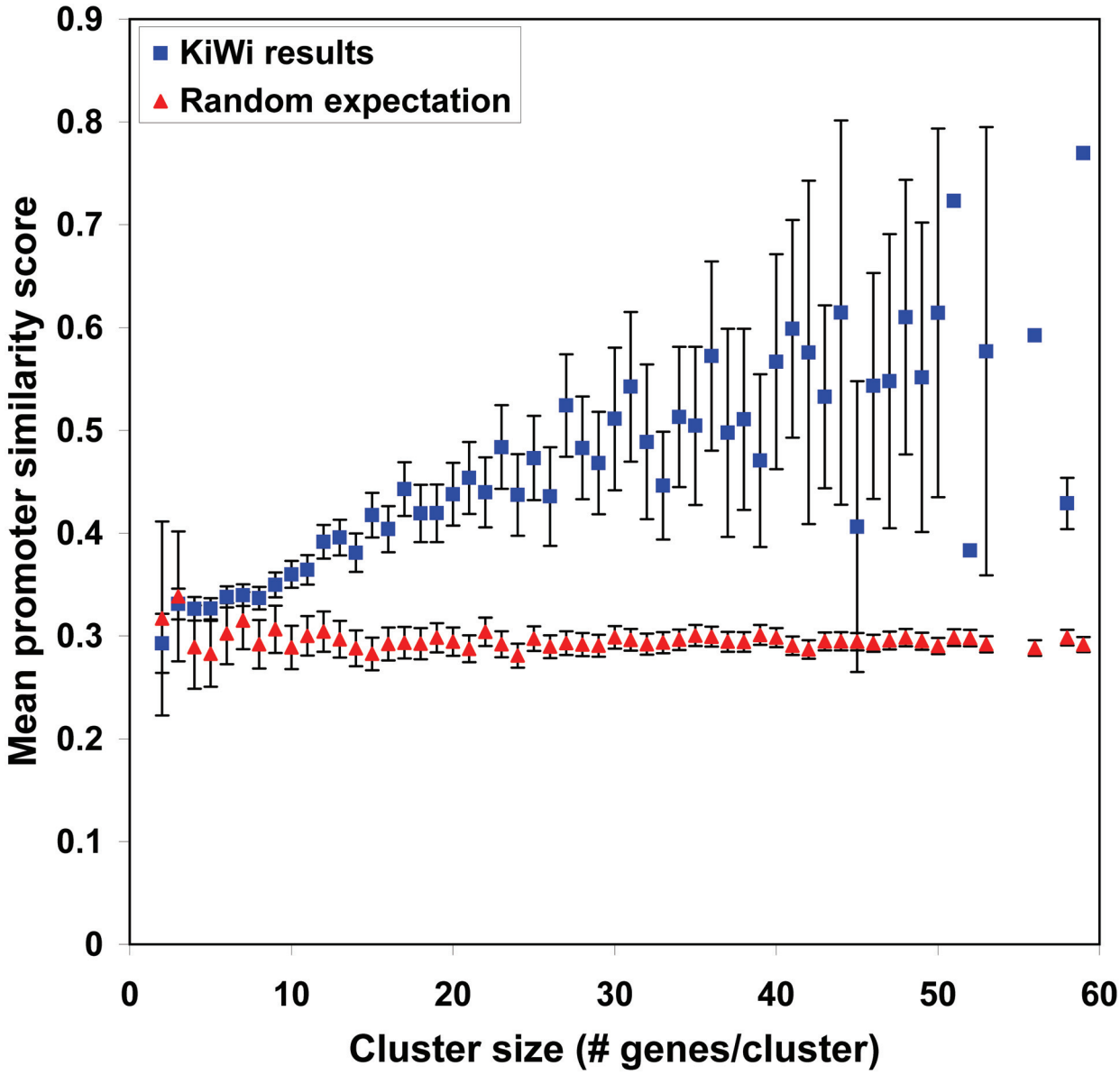


Figure 3.12. cisRED analysis for Cooper promoter dataset comparing different pattern lengths

The mean promoter similarity score for each pattern length is shown. As with cluster size (Figure 3.11) the promoter similarity score increases with greater pattern length (number of experiments). Error bars indicate 95% confidence limits.

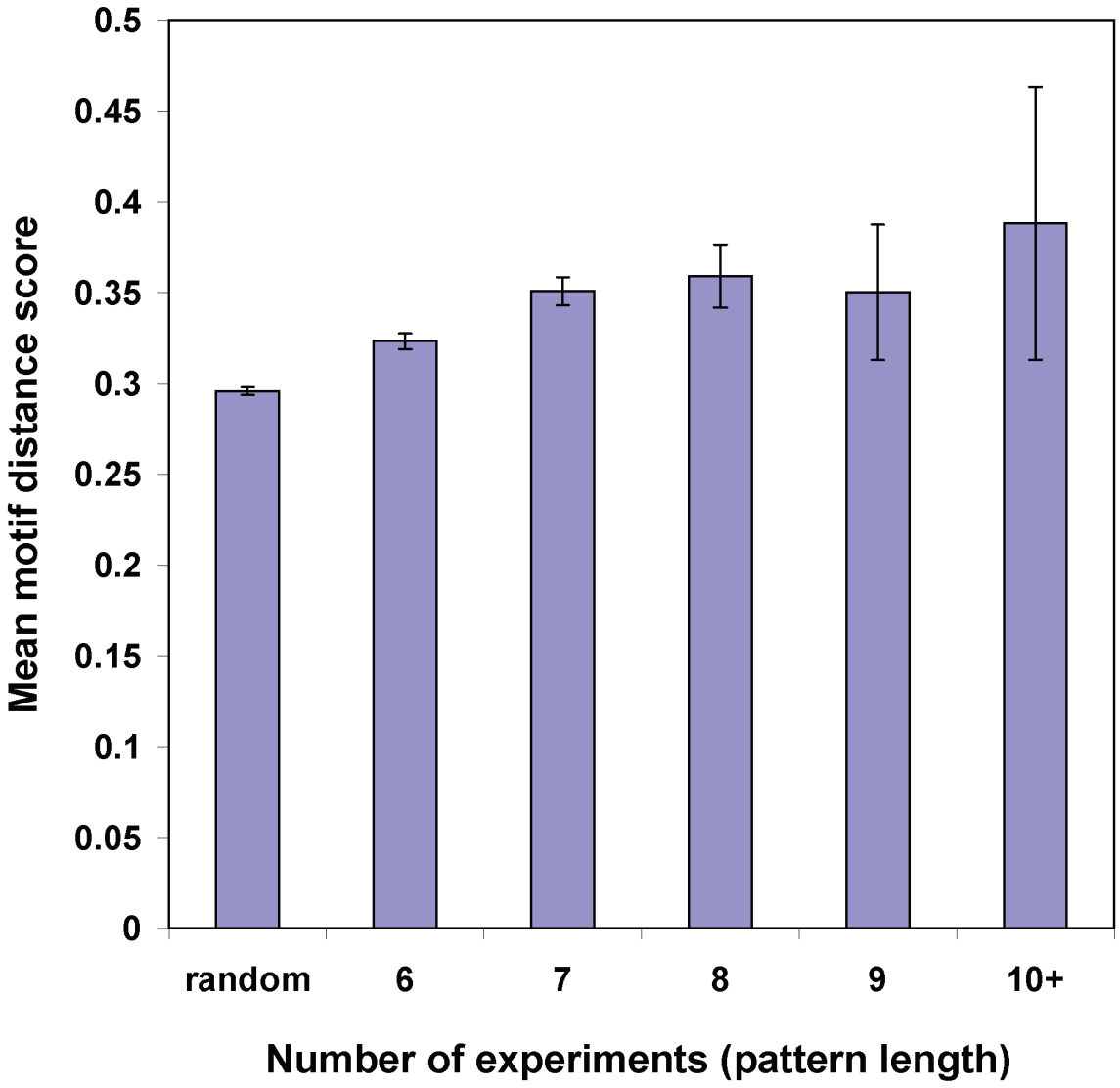
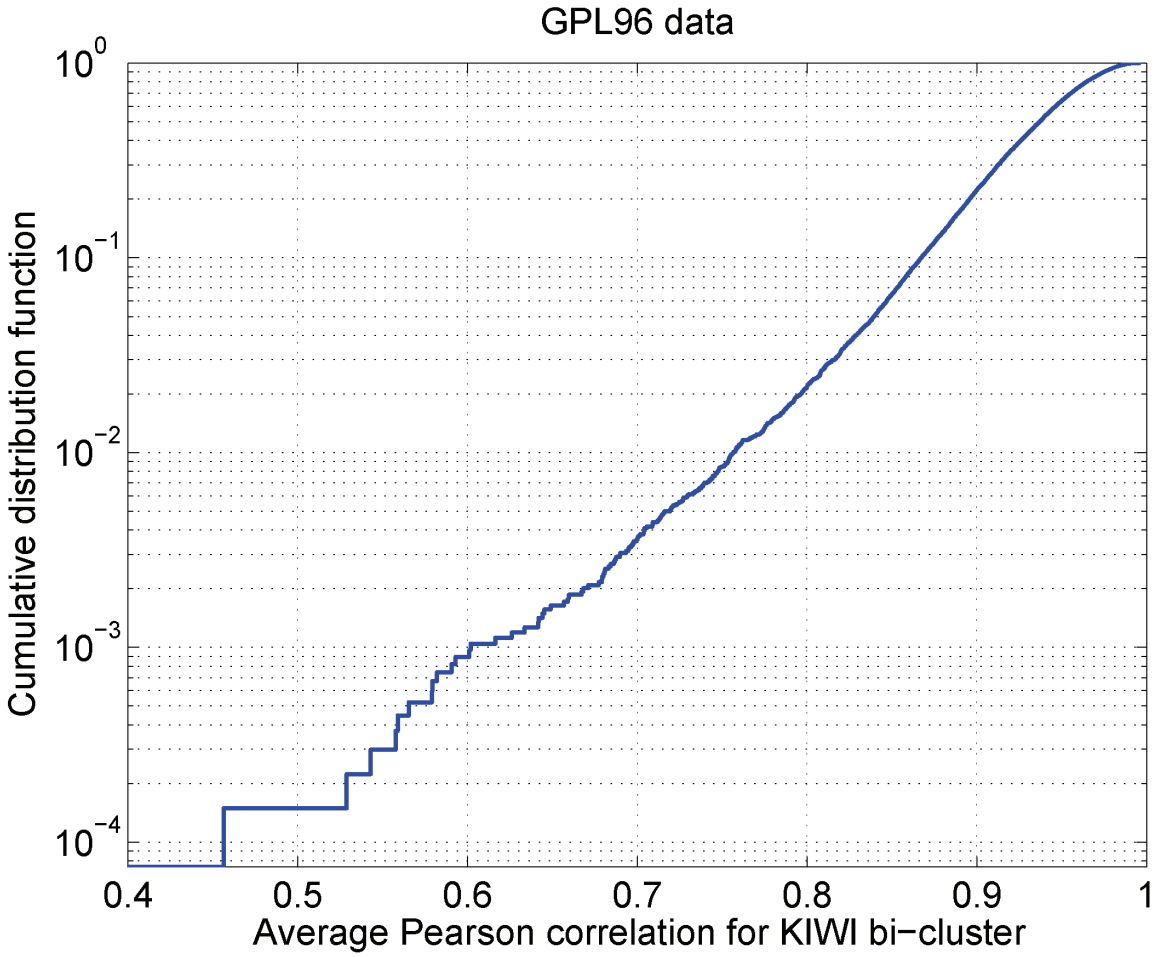


Figure 3.13. Distribution of mean Pearson correlations for all KiWi clusters for GPL96 data

The figure shows that the vast majority of subspace clusters (bi-clusters) have very high average pairwise correlations with 90% of clusters having an average r > 0.95.



References

1. Jensen, L.J., J. Lagarde, C. von Mering, and P. Bork, *ArrayProspector: a web resource of functional associations inferred from microarray expression data.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W445-8.
2. Segal, E., N. Friedman, D. Koller, and A. Regev, *A module map showing conditional activity of expression modules in cancer.* Nat Genet, 2004. **36**(10): p. 1090-8.
3. Oldham, M.C., S. Horvath, and D.H. Geschwind, *Conservation and evolution of gene coexpression networks in human and chimpanzee brains.* Proc Natl Acad Sci U S A, 2006. **103**(47): p. 17973-8.
4. Stuart, J.M., E. Segal, D. Koller, and S.K. Kim, *A gene-coexpression network for global discovery of conserved genetic modules.* Science, 2003. **302**(5643): p. 249-55.
5. Kemmeren, P., N.L. van Berkum, J. Vilo, T. Bijma, R. Donders, A. Brazma, and F.C. Holstege, *Protein interaction verification and functional annotation by integrated analysis of genome-scale data.* Mol Cell, 2002. **9**(5): p. 1133-43.
6. Walhout, A.J., J. Reboul, O. Shtanko, N. Bertin, P. Vaglio, H. Ge, H. Lee, L. Doucette-Stamm, K.C. Gunsalus, A.J. Schetter, D.G. Morton, K.J. Kemphues, V. Reinke, S.K. Kim, F. Piano, and M. Vidal, *Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline.* Curr Biol, 2002. **12**(22): p. 1952-8.
7. Haverty, P.M., U. Hansen, and Z. Weng, *Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification.* Nucleic Acids Res, 2004. **32**(1): p. 179-88.
8. Segal, E., H. Wang, and D. Koller, *Discovering molecular pathways from protein interaction and gene expression data.* Bioinformatics, 2003. **19 Suppl 1**: p. i264-71.
9. Ge, H., A.J. Walhout, and M. Vidal, *Integrating 'omic' information: a bridge between genomics and systems biology.* Trends Genet, 2003. **19**(10): p. 551-60.
10. Li, S., C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.O. Vidalain, J.D. Han, A. Chesneau, T. Hao, D.S. Goldberg, N. Li, M. Martinez, J.F. Rual, P. Lamesch, L. Xu, M. Tewari, S.L. Wong, L.V. Zhang, G.F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H.W. Gabel, A. Elewa, B. Baumgartner, D.J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S.E. Mango, W.M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K.C. Gunsalus, J.W. Harper, M.E. Cusick, F.P. Roth, D.E. Hill, and M. Vidal, *A map of the interactome network of the metazoan C. elegans.* Science, 2004. **303**(5657): p. 540-3.
11. Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet, 2003. **34**(2): p. 166-76.
12. Ihmels, J., S. Bergmann, and N. Barkai, *Defining transcription modules using large-scale gene expression data.* Bioinformatics, 2004. **20**(13): p. 1993-2003.
13. Leyfer, D. and Z. Weng, *Genome-wide decoding of hierarchical modular structure of transcriptional regulation by cis-element and expression clustering.* Bioinformatics, 2005. **21 Suppl 2**: p. ii197-ii203.
14. Haberer, G., M.T. Mader, P. Kosarev, M. Spannagl, L. Yang, and K.F. Mayer, *Large-scale cis-element detection by analysis of correlated expression and sequence conservation between Arabidopsis and Brassica oleracea.* Plant Physiol, 2006. **142**(4): p. 1589-602.
15. Ben-Dor, A., B. Chor, R. Karp, and Z. Yakhini, *Discovering local structure in gene expression data: the order-preserving submatrix problem.* J Comput Biol, 2003. **10**(3-4): p. 373-84.

16. De Smet, F., J. Mathys, K. Marchal, G. Thijs, B. De Moor, and Y. Moreau, *Adaptive quality-based clustering of gene expression profiles.* Bioinformatics, 2002. **18**(5): p. 735-46.
17. Prelic, A., S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, *A systematic comparison and evaluation of biclustering methods for gene expression data.* Bioinformatics, 2006. **22**(9): p. 1122-9.
18. Wu, C.J. and S. Kasif, *GEMS: a web server for biclustering analysis of expression data.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W596-9.
19. Madeira, S.C. and A.L. Oliveira, *Biclustering algorithms for biological data analysis: a survey.* IEEE/ACM Trans Comput Biol Bioinform, 2004. **1**(1): p. 24-45.
20. Liu, J. and W. Wang, *Op-cluster: Clustering by tendency in high dimensional space*, in *Proceedings of the 3rd IEEE International Conference on Data Mining*. 2003, IEEE Computer Society: Melbourne, FL, USA. p. 187-194.
21. Pavuna, D., *Modern physics in a global society.* Fizika A, 1999. **8**(4): p. 205-14.
22. Barrett, T., T.O. Suzek, D.B. Troup, S.E. Wilhite, W.C. Ngau, P. Ledoux, D. Rudnev, A.E. Lash, W. Fujibuchi, and R. Edgar, *NCBI GEO: mining millions of expression profiles--database and tools.* Nucleic Acids Res, 2005. **33**(Database issue): p. D562-6.
23. Sherlock, G., T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J.C. Matese, S.S. Dwight, M. Kaloper, S. Weng, H. Jin, C.A. Ball, M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, and J.M. Cherry, *The Stanford Microarray Database.* Nucleic Acids Res, 2001. **29**(1): p. 152-5.
24. Pei, J., J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.C. Hsu, *PrefixSpan: Mining Sequential Patterns Efficiently by Prefix Projected Pattern Growth*, in *Proceedings of the 2001 International Conference on Data Engineering*. 2001, IEEE Educational Activities Department: Heidelberg, Germany. p. 215-26.
25. Gao, B.J., O.L. Griffith, M. Ester, and S.J. Jones, *Discovering significant OPSM subspace clusters in massive gene expression data*, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2006, ACM Press: Philadelphia, PA, USA. p. 922-8.
26. Agrawal, R. and R. Srikant, *Mining Sequential Patterns*, in *Proceedings of the Eleventh International Conference on Data Engineering*. 1995, IEEE Computer Society: Taipei, Taiwan. p. 3-14.
27. Dhillon, I.S., E.M. Marcotte, and U. Roshan, *Diametrical clustering for identifying anti-correlated gene clusters.* Bioinformatics, 2003. **19**(13): p. 1612-9.
28. Qian, J., M. Dolled-Filhart, J. Lin, H. Yu, and M. Gerstein, *Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions.* J Mol Biol, 2001. **314**(5): p. 1053-66.
29. Cooper, S.J., N.D. Trinklein, E.D. Anton, L. Nguyen, and R.M. Myers, *Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.* Genome Res, 2006. **16**(1): p. 1-10.
30. Gentleman, R.C., V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y. Yang, and J. Zhang, *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol, 2004. **5**(10): p. R80.
31. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock,

Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000. **25**(1): p. 25-9.
32. Zeeberg, B.R., H. Qin, S. Narasimhan, M. Sunshine, H. Cao, D.W. Kane, M. Reimers, R.M. Stephens, D. Bryant, S.K. Burt, E. Elnekave, D.M. Hari, T.A. Wynn, C. Cunningham-Rundles, D.M. Stewart, D. Nelson, and J.N. Weinstein, *High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID).* BMC Bioinformatics, 2005. **6**: p. 168.
33. Ho Sui, S.J., J.R. Mortimer, D.J. Arenillas, J. Brumm, C.J. Walsh, B.P. Kennedy, and W.W. Wasserman, *oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes.* Nucleic Acids Res, 2005. **33**(10): p. 3154-64.
34. Robertson, G., M. Bilenky, K. Lin, A. He, W. Yuen, M. Dagpinar, R. Varhol, K. Teague, O.L. Griffith, X. Zhang, Y. Pan, M. Hassel, M.C. Sleumer, W. Pan, E.D. Pleasance, M. Chuang, H. Hao, Y.Y. Li, N. Robertson, C. Fjell, B. Li, S.B. Montgomery, T. Astakhova, J. Zhou, J. Sander, A.S. Siddiqui, and S.J. Jones, *cisRED: a database system for genome-scale computational discovery of regulatory elements.* Nucleic Acids Res, 2006. **34**(Database issue): p. D68-73.
35. Griffith, O.L., E.D. Pleasance, D.L. Fulton, M. Oveisi, M. Ester, A.S. Siddiqui, and S.J. Jones, *Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses.* Genomics, 2005. **86**(4): p. 476-88.
36. Allocco, D.J., I.S. Kohane, and A.J. Butte, *Quantifying the relationship between co-expression, co-regulation and gene function.* BMC Bioinformatics, 2004. **5**: p. 18.

4. Meta-analysis and tissue microarray analysis identifies important diagnostic and prognostic biomarkers in thyroid cancer[8,9,10]

4.1. Introduction

Thyroid nodules are extremely common, being palpable in 4-7% of the North American adult population, with new nodules detected at a yearly rate of 0.1% [1, 2]. Currently, fine needle aspiration biopsy (FNAB) represents the most important initial test for diagnosing malignancy. The result of the FNAB cytology can be classified as: benign (70% of cases), malignant (5-10%), indeterminate/suspicious (10-20%), or non-diagnostic (10-15%) [3-5]. While non-diagnostic FNABs can be repeated, the indeterminate or suspicious group presents a dilemma for the clinician. In a recent report on 80 patients who underwent thyroid resection for an indeterminate FNAB diagnosis of follicular neoplasm (FN) only 20% were confirmed as malignant [6]. Thus, many patients undergo thyroid surgery for nodular disease which is eventually diagnosed as benign disease. Therefore, one of the main challenges in thyroid cancer is the correct diagnosis of malignant from benign in cases where the FNAB fails.

Given the diagnostic limitations of FNAB when applied to thyroid tumours, multiple investigators have carried out expression profiling studies with hopes of identifying new diagnostic tools. Such analyses attempt to identify differentially expressed genes with an important role in disease development or progression using large-scale transcript-level expression profiling technologies such as cDNA microarrays [7], oligonucleotide arrays [8] and Serial Analysis of Gene Expression (SAGE) [9]. Typically, dozens or hundreds of genes are identified, many of which are expected to be false-positives and only a small fraction useful as diagnostic/prognostic markers or therapeutic targets. A logical approach to distinguishing important genes from spurious genes, given a large number of candidate gene lists is to search

[8] A portion of this chapter has been published. Griffith OL, Melck A, Jones SJM, Wiseman SM. 2006. A Meta-analysis and Meta-review of Thyroid Cancer Gene Expression Profiling Studies Identifies Important Diagnostic Biomarkers. Journal of Clinical Oncology. 24(31):5043-505.

[9] A portion of this chapter has been published. Wiseman SM, Griffith OL, Deen S, Rajput A, Masoudi H, Gilks B, Goldstein L, Gown A, Jones SJM. 2007. Identification of Molecular Markers Altered During Transformation of Differentiated Into Anaplastic Thyroid Carcinoma. Arch Surg. 142(8):717-729.

[10] Co-authorship details: I was the primary author of Griffith *et al*. (2006). I performed the bulk of the statistical analysis and contributed text and figures to Wiseman *et al*. (2007). I was responsible for all analysis, text, figures and tables included in this chapter except where indicated below. The tissue microarray design, construction, staining, and scoring were performed by Sam Wiseman and members of his laboratory (see sections 4.2.2.1 and 4.2.2.2). The clinical and pathological outcome data and ethics submissions were also performed by Sam Wiseman and laboratory members. Sam Wiseman contributed to the writing of sections 4.1, 4.2.2, 4.4.2.2, and 4.5 and provided Tables 4.3-4.5. Adrienne Melck contributed to the writing of 4.1, 4.4.1.2 and 4.4.1.3. Sam Wiseman and Steven Jones funded and supervised the project.

for the intersection of genes identified in multiple independent studies [10]. It is expected that biologically relevant genes will be over-represented and system-specific spurious genes under-represented. As large numbers of cancer profiling studies have become available the identification of such intersections has become increasingly popular [10-12] but none have investigated thyroid cancer specifically. Such studies, while conceptually simple, face a number of technical challenges such as: inconsistent gene identifiers; inaccessible data; and uncertain significance of results. Here, we attempt to overcome these challenges.

Our approach involves a 'vote-counting' strategy based on the number of studies reporting a gene as differentially expressed and further ranking based on total sample size and average fold-change. This strategy is similar to the approach we used to show that gene pairs consistently coexpressed in multiple platforms are more likely to share a common biological process (chapter 2). In this chapter, our primary objective was to use validation from multiple expression profiling datasets to identify high-confidence differentially expressed genes as potential biomarkers for distinguishing benign from malignant thyroid cancer. We present a novel 'meta-review' method for ranking genes based on published evidence, successfully validate our method against a more traditional meta-analysis approach and provide a large number of highly significant 'multi-study' genes. Such markers should prove a useful resource for further study by high throughput molecular analytic techniques. We also present preliminary results for a tissue microarray (TMA) analysis of clinical samples in which 100 benign thyroid samples were compared to 96 malignant thyroid carcinoma samples for staining of 56 antibodies. A subset of the antibodies target genes that were predicted through the meta-analysis and the rest were chosen based on availability or other criteria. We identified a number of markers with significant differences and attempted to develop a panel of markers with diagnostic classification utility. Genes identified in the meta-analysis were in all cases found to be significantly differentially expressed on the TMA and in general, higher rank in the meta-analysis correlated with better classification performance.

Another challenge in thyroid cancer treatment is the development of an understanding of aggressive tumour characteristics or transformation. Transformation is a term commonly used to describe the biological process in which normal or premalignant cells undergo a change and become cancerous [13]. For thyroid cancer, transformation or anaplastic transformation describes an intratumoural evolution, or progression, from differentiated thyroid carcinoma (DTC) into

anaplastic thyroid carcinoma (ATC) [14]. DTC, which includes papillary thyroid carcinoma (PTC) and follicular thyroid carcinoma (FTC), accounts for more than 90% of newly diagnosed thyroid malignancies, and with current treatment few individuals die of their disease [15]. Even individuals who present with evidence of metastatic DTC can experience long-term survival [16]. In contrast, ATC accounts for less than 2% of all newly diagnosed thyroid malignancies but is considered one of the most lethal of all human cancers [17]. The rarity and rapidly fatal disease course of ATC have made the study of this cancer difficult in both the clinic and laboratory. Accumulated clinical, pathologic, and experimental evidence has led to acceptance of the hypothesis that ATC transforms or evolves from pre-existing DTC [14]. However, little understanding exists of the specific molecular alterations and mechanisms that underlie this evolution and only a couple expression profiling studies with relatively small sample sizes have been carried out [18, 19].

To develop a better understanding of thyroid tumour progression, the second objective of this chapter was to evaluate the change in the tumour expression profile that occurs during the transformation of DTC into ATC. A second TMA analysis was performed in which 12 undifferentiated ATC samples were compared to 12 patient-matched DTC samples for staining of 62 antibodies. The 62 molecular markers were chosen because they represent a wide variety of gene products important in the maintenance of normal cellular function. Some of these proteins are thyroid specific, others have previously been identified as altered by anaplastic transformation, and yet others have never been well characterized in a thyroid cancer cohort. A number of significant markers were identified which could help to improve our understanding of the anaplastic transformation process.

4.2. Methods

4.2.1. Meta-analysis

4.2.1.1. Data collection and curation

Lists of differentially expressed genes were collected and curated from publications, supplementary web pages, or files provided by authors. A total of 34 comparisons were available from 21 studies, utilizing 10 different expression platforms (Table 4.1). The following information was recorded wherever possible: Unique identifier (probe/tag/accession); gene name/description; gene symbol; comparison conditions; sample numbers for each condition; fold change; direction of change; and Pubmed ID. For consistency, all fold change values were

converted to whole number fold-changes with +/- sign indicating direction of change. For example a 0.5 fold change for cond1/cond2 was converted to -2.0. If 'Signal Log Ratio' (SLR) was provided this was also converted (2^{SLR} = fold change). Individual data tables were stored in separate files and then combined into a single master file. In one study, Giordano *et al.* (2005) were interested in identifying expression profiles that correlated with specific mutational status (e.g. BRAF V600E point mutation, RET/PTC rearrangement, etc) [20]. The comparisons conducted were unique to this study and not applicable for our overlap analysis. Therefore we re-analyzed their data using the log normalized data file (provided as supplementary data) to calculate differentially expressed genes between cancer (PTC) and non-cancer (Normal) samples. A total of 90 up-regulated and 151 down-regulated genes were chosen with a fold-change of greater than 2.0 and Bonferroni corrected p-value (t-test, two-sided) of less than 0.001. Note: all abbreviations used for sample descriptions are explained in Table 4.2.

4.2.1.2. Gene mapping

In order to determine the amount of overlap between the various published studies, it was first necessary to obtain consistent gene identifiers. Entrez gene identifiers from NCBI were chosen as the target identifier. SAGE tags were mapped to genes by the first position (3'-most NlaIII anchoring enzyme recognition site), sense-strand tag predicted from Refseq [21] or MGC [22] sequences and then mapped to Entrez using the DiscoverySpace software package (http://www.bcgsc.ca/discoveryspace/) [23]. Only unambiguous mappings were allowed. Affymetrix probes were mapped to Entrez gene ids using the Affymetrix annotation files (http://www.affymetrix.com). Clone accession ids were mapped to Entrez gene ids using the DAVID Resource (http://david.abcc.ncifcrf.gov/) [24]. If no tag, probe, or accession id was available, the entry was mapped from gene symbol directly to Entrez or indirectly using gene synonyms. In 43 cases, Entrez gene ids were manually determined based on gene description. Any genes still not mapped were assigned an Entrez id of 'NA'. Gene history was checked to identify Entrez ids that have been retired or replaced. If retired, Entrez id was changed to 'NA'. If replaced, the new replacement Entrez id was used.

4.2.1.3. Ranking

Each published study consists of one or more comparisons between a pair of conditions (e.g. PTC vs. normal) resulting in a list of differentially expressed genes. A method of ranking potential molecular markers was devised for each comparison group. A comparison group refers

to a list of comparisons that address a common question of interest. For example, to identify markers that consistently distinguish cancer from non-cancer (normal or benign) we would analyze all the comparisons that contrast 'cancer' samples (i.e., PTC, FTC, ATC, etc) against 'non-cancer' samples (i.e., Norm, GT, FA, etc). It should be noted that two studies by Finley and colleagues [25, 26] appear to have a high amount of redundancy between the samples analyzed for gene expression. Therefore, to prevent artificially high overlap between these datasets, only one or the other was included in each comparison group (depending on the nature of the comparison).

Genes were ranked according to several criteria in the following order of importance: (1) Number of comparisons in agreement (i.e. listing the same gene as differentially expressed and with a consistent direction of change); (2) Total number of samples for comparisons in agreement; and (3) Average fold change reported for comparisons in agreement. Total sample size was considered more important than average fold change because many studies do not report a fold change. Therefore, average fold-change was based solely on the subset of studies for which a fold-change value was available. Also, it should be noted that the average fold-change is based on all reported fold-changes irrespective of the method used to calculate them (normalization, logging, etc). In most cases fold changes are calculated as a ratio of mean expression in one condition versus another. But, in other cases where matched normal and tumour patient samples were available it could represent the median fold change observed for all individual tumour/normal ratios [27]. Despite this caveat, we believe that the average and range of reported fold changes gives a good idea of the relative magnitude of differential expression for each gene of interest.

4.2.1.4. Assessment of significance

Significance of the observed level of overlap between studies for each comparison subset was assessed by Monte Carlo simulation using custom Perl scripts. Where possible, the actual gene lists produced by mapping each expression technology to Entrez gene ID were utilized. For studies with custom arrays [18, 28-30] the appropriate number of genes was chosen from the combined gene list of all other platforms. Since the custom arrays may in reality have features/genes unique to the array, this simplifying step may actually increase the chance of observing overlap in the random simulations. Thus, the final p-value is likely an overestimate. For most platforms, mapping of features to genes was not perfect. On average, 91.4% of features

were successfully mapped to a gene identifier. Thus, for the custom arrays, we randomly chose this same fraction of genes. For SAGE, three thyroid libraries from CGAP[31] (normal, benign, and carcinoma) were used to create a realistic total tag set and then mapped to Entrez as above. Once total gene lists were created for each platform type, we randomly created gene subsets of the same size observed in our review of the literature. For example, in the 'cancer vs. non-cancer' analysis, one comparison (PTC vs. Norm) identified 24 up- and 27 down-regulated genes with the Affymetrix HG-U95A platform[32]. In our simulation, we would randomly select and label 24 'up' and 27 'down' genes from the Affymetrix HG-U95A total gene list. A similar random selection was performed for all other comparisons in the 'cancer vs. non-cancer' subset using the appropriate total gene lists. Finally, the amount of overlap between comparisons was tallied as in the real analysis. This entire process was repeated 10,000 times to produce a distribution of overlap results from the random simulations. A p-value was then estimated by comparing the actual overlap result to the random distribution. A result was considered significant at $p<0.05$.

4.2.1.5. Meta-analysis of raw Affymetrix data

The method presented above makes use of reported lists of differentially expressed genes from published literature. An obvious disadvantage of this approach is that each publication may make use of different methods to ascertain differential expression (scaling, filtering, normalization, significance thresholds, p-value estimation, multiple testing corrections, etc.). Collecting and re-analyzing 21 sets of raw data from 10 different platforms in a consistent manner would be an immense task and most likely impossible, because many raw datasets are unavailable. The majority of data had not been placed in public databases. However, to assess our method, we did re-analyze a subset of data from raw image files using a standard methodology. The Affymetrix platforms were chosen for their ease of analysis and because they were most represented in the published studies (nine studies). Two of the datasets were freely available on the web [20, 32] and the other seven were requested by email. Ultimately, we were able to obtain only four datasets (two requests were successful) representing five comparisons with a total of 117 samples (65 PTC, 15 FTC, 25 normal and 12 FA) [20, 32-34]. Cel files were loaded and analyzed with the DChip software (Build date: Nov 17 2005). Arrays were normalized and modeled using default settings. Probes were filtered based on variation (0.50 < Standard deviation / Mean < 1,000.00) and P call across samples (>= 20% of arrays). Probes were determined to be differentially expressed if they demonstrated fold change greater than two

and p-value less than 0.05 (after FDR-based multiple testing correction). Finally, the five comparisons (3 PTC vs. Norm; 1 FTC vs. Norm; and 1 FTC vs. FA) were analyzed for overlapping genes as above and the results compared to the cancer versus non-cancer comparison analysis for concordance using the LOLA tool [11].

4.2.1.6. Gene ontology analysis

A Gene Ontology [35] analysis was performed for the genes with multi-study confirmation in the cancer versus non-cancer overlap analysis group using the BINGO [36] plug-in for the Cytoscape [37] software package. Significance was calculated using the hypergeometric test, corrected with a Benjamini & Hochberg False Discovery Rate (FDR) correction, and a cut off of 0.05 applied to the result.

4.2.1.7. Supplementary materials

Supplementary data files related to the meta-analysis can be obtained at www.bcgsc.ca/bioinfo/ge/thyroid/.

4.2.2. Tissue microarray

4.2.2.1. Study designs

4.2.2.1.1. Malignant versus benign

The TMA was composed of 100 benign thyroid lesions (54 GT, 26 FA, 10 HCA, 4 HN, 3 HT, and 3 LT) and 105 malignant lesions (90 PTC, 6 FTC, 6 MTC, and 3 HCC). Refer to Table 4.2 for explanation of thyroid sample abbreviations. The 205 study subjects were selected from a Thyroid Surgery Patient Database that is prospectively maintained by Dr. Sam Wiseman. All study subjects had undergone thyroid surgery for benign or malignant disease between January 2001 and May 2005 and the medical records of these individuals were retrospectively reviewed for: patient demographics, surgical procedure, pathologic diagnosis, extrathyroidal extension by cancer, vascular invasion by cancer, multifocality of cancer, completeness of cancer resection, presence of lymph node and/or distant metastases, adjuvant therapy with either radioactive iodine (RAI) and/or external beam radiation therapy (EBRT), postoperative thyroglobulin levels, AJCC cancer stage, AMES scores, length of postoperative follow-up, and disease status at last physician follow up visit. Subsequently we obtained the archival pathology specimens of these patient's thyroid tumours for TMA construction. This study was carried out with the approval of our Institutional Research Ethics Boards. In order to ensure a differentiated thyroid cancer

(DTC) cohort for the study cancer population, the primary cancers diagnosed as either medullary thyroid carcinoma (MTC) or Hurthle cell carcinomas (HCC) were excluded from the statistical analysis. Therefore, the final study cohort was composed of 100 individuals diagnosed with benign thyroid lesions and 96 individuals diagnosed with either papillary or follicular carcinoma.

4.2.2.1.2. ATC versus DTC

Sequential archival cases of ATC with an adjacent associated DTC focus, and with available paraffin blocks, that had been diagnosed and treated in British Columbia during a 20-year period (January 1, 1984, through December 31, 2004) were identified through the provincial tumour registry for tissue microarray (TMA) construction. The study was approved by the research ethics boards of the University of British Columbia, British Columbia Cancer Agency, Vancouver Coastal Health, and Providence Health Care (Vancouver, British Columbia). All patients had newly diagnosed ATC, and all clinical data were retrospectively collected from hospital medical records. Clinicopathologic data collected included patient age, patient sex, type of therapy, patient follow-up, and survival. Haematoxylin and Eosin stained sections of each tumour were examined, and areas of ATC or DTC (PTC or FTC) were marked on both the slide and the corresponding paraffin block for TMA construction. Adequate tissue was present for immunohistochemical staining of 12 ATC and 12 associated adjacent DTC foci.

4.2.2.2. Tissue microarray construction, staining, and evaluation

Archival paraffin-embedded tissue specimens from the study patients were reviewed by pathologist to confirm the diagnosis and identify an area of the specimen that would be incorporated into the TMA. TMA construction was carried out utilizing a tissue-arraying instrument (Beecher Instruments, Silver Springs, MD) which transferred two 0.6-mm tissue cores from the pre-selected regions of each primary block to defined array coordinates in the recipient TMA blocks. A Leica microtome was used to cut serial 4-μm sections from the TMA blocks that were then transferred onto adhesive-coated glass slides for staining. Sections were then deparaffinised and antigen retrieval was carried out. Antibodies were optimized for thyroid tissue according to the manufacturer's instructions and appropriate positive and negative controls were used for each antibody. Pathologists, blinded to all clinical data, examined the stained TMA sections at high power magnification in order to determine the proportion of cells expressing the markers. Any inter-pathologist discrepancy in the scoring of a specific tissue core was immediately resolved. If the scores for the two samples from each specimen were different, the

higher of the two was assigned for analysis. The 65 antibodies used, their target genes, cellular localization, scoring system and the arrays probed are summarized in Table 4.3. Additional antibody characteristics such as the isotype, clone, company, catalogue number, antigen retrieval method, and concentration are summarized in Table 4.4. In total, 56 antibodies were processed for the benign/malignant array and 62 for the ATC/DTC array. The scoring systems utilized were based on previously published reports of immunohistochemical studies evaluating these markers, and are summarized in Table 4.5. All scores were recorded in a standardized TMA case map that corresponded to each TMA section (Microsoft Excel; Microsoft, Redmond, WA). All data were processed by custom TMA-deconvoluter software (developed using the Perl programming language). The deconvoluted data was then transferred into a master database, which included all clinical and pathologic data, for statistical analysis.

4.2.2.3. Statistical analysis

Significant associations between marker staining and pathologic status (ATC/DTC or malignant/benign) were determined using contingency table statistics (Pearson χ^2 or Fisher exact test where appropriate). For both arrays, two marker score groupings were analyzed. In the first grouping, marker scores were grouped as either negative (score=0) or positive (score≥1). In the second grouping, marker scores were grouped as either negative/low (score≤1) or medium/high (score≥2). A marginal homogeneity (MH) test for 2 related samples was used to test for a significant trend toward increasing or decreasing score for a marker between the matched pairs of ATC and DTC samples for the anaplastic array. A Mann-Whitney U-test (MU) test for two independent samples was used to test for a significant trend towards increasing or decreasing score for a marker between the benign and malignant samples. The MH and MU tests do not require the marker scores to be grouped but instead uses the actual semi-quantitative scores (ungrouped). All statistics were corrected for multiple testing using the Benjamini and Hochberg (BH) correction [38]. The BH correction is a simple step-up, false-discovery rate–controlling procedure that is much less stringent than the more commonly used Bonferroni correction. All statistical tests were 2-sided and considered statistically significant at $p<0.05$ (after correction). All statistics were performed with the SPSS statistical software package (version 13.0; SPSS Inc, Chicago, Illinois) or the R statistical programming language (version 2.3.1; R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria)

The samples and markers were clustered using a simple hierarchical clustering algorithm and heat maps generated to visualize the data using the R 'gplots' library (version 2.3.0). The markers were also evaluated for their utility in classification (ATC versus DTC or malignant versus benign) using the Random Forests classifier algorithm (Random Forests, version 1.0; Salford Systems, San Diego, California).

4.3. Results

4.3.1. Meta-analysis

A total of 34 comparisons were available from 21 studies, utilizing 10 different expression platforms (Table 4.1). Of the 1,785 genes reported as differentially expressed in these studies (827 up-, 958 down-regulated), 1,562 could be mapped to an Entrez gene identifier (723 up-, 839 down-regulated). In all overlap analysis groups considered except for one, we identified genes that were reported in multiple studies with a level of overlap found to be significant by Monte Carlo simulation ($p<0.05$) (Table 4.6). The 'cancer versus non-cancer' group is provided as the most relevant example in the following results and discussion. In this case, a total of 755 genes were reported from 21 comparisons, and of these, 107 genes were reported more than once with a consistent fold-change direction (Figure 4.1). In three cases (MET, TFF3, and SERPINA1), genes were independently reported as many as 6 times. Only 18 genes were found to be reported in multiple studies with inconsistent fold-change. This in itself is an encouraging result. Given that approximately equal numbers of genes were reported as up- versus down-regulated (723 up, 839 down) we might expect that multi-study genes with inconsistent fold change direction would be as (or more) common than genes with consistent direction (under random expectation). Instead, we see that in most cases (85.6%), studies that report the same gene agree on the direction, even for large numbers of studies (where the chance of a spurious discrepancy is increased).

The total amount of overlap observed was assessed by Monte Carlo simulation. Real data was found to have significantly more overlap than simulated data ($p<0.0001$, 10,000 permutations). In the simulation, an average of 18.2 (95% CI, 18.12 to 18.28) genes were observed with an overlap of two (same gene identified in two comparisons) compared to 68 in the actual data. For overlap of three, only 0.3 (95% CI, 0.29 to 0.31) genes were observed on average compared to 27 for real data. In 10,000 permutations, the simulated data never produced an overlap greater than three whereas real data identified 12 genes with overlap of four, five or six. The probability

of observing one or more genes with an overlap of two or more was p=1.0. For overlap of three or more p=0.037, and for four or more p<0.0001. The total number of genes with overlap of two was still highly significant but we expect at least some false-positives to occur by chance in this category. Therefore, we have provided only those genes (top 39) with overlap of three or more and consider those with four or more to be the most reliable (Table 4.7). The complete list is available as supplementary data on our supplementary web page (http://www.bcgsc.ca/platform/bioinfo/ge/thyroid/).

If the 'cancer vs. non-cancer' group is broken into two categories, 'cancer vs. normal' and 'cancer vs. benign', we see that most of the top genes were found in both types of comparisons. A small number of genes were found in only one of the two categories (cancer vs. normal, cancer vs. benign, or both). In all, 58.9% (63/107) of the multi-study genes (two or more overlapping studies) were found in both a cancer versus normal and cancer versus benign comparison (Figure 4.2).

A comparison of genes with multi-study evidence based on published lists versus the smaller subset re-analyzed from raw Affymetrix microarray data showed a highly significant level of agreement (p<0.0001). The 107 'cancer versus non-cancer' multi-study genes showed a concordance of 0.177 (95% CI, 0.129 to 0.225) with the 179 multi-study genes identified from the re-analyzed Affymetrix subset (Figure 4.3). In total, there were 43 genes identified by both methods. Given that the two lists of genes were produced by very different subsets of data, in addition to the potential differences in processing, this was an encouraging result. However, it does appear that re-processing the microarray data in a consistent manner would alter the results and would likely increase the total number of multi-study genes.

A gene ontology analysis of multi-study genes from the cancer versus non-cancer overlap analysis group identified 12 significant terms (Table 4.8). Genes tended to be localized in the extracellular region and more specifically the extracellular matrix. Significant biological processes involved hormone metabolism and generation. Significant molecular functions involved binding (cadmium, selenium, and copper), MAP kinase phosphatase activity and retinoic acid receptor activity.

4.3.2. Tissue microarray

4.3.2.1. Malignant versus benign

A large number of the 56 markers tested showed significant associations with diagnosis (Table 4.9 and Table 4.10). There were 30 significant markers in grouping 1, 15 in grouping 2 and 33 in the ungrouped analysis. In total, 33 markers were found to be significantly associated in one or more of the three tests after multiple testing correction. Of these, 7 markers were down-regulated and 26 up-regulated (in malignant compared to benign). The Random Forests algorithm was able to achieve a good classification of patients into their correct diagnostic group using marker score. Specifically, cross-validation analysis reported an ROC integral of 0.971, an overall accuracy of 91.7%, sensitivity of 87.5% and specificity of 96%. This translates to a misclassification of only 4 out of 100 benign and 12 of 96 malignant samples. This performance is graphically illustrated (Figure 4.4 and Figure 4.5) by the good separation of benign (indicated by the light green side bar) from malignant (dark green side bar) samples in the hierarchical clustering heat maps. A number of variables contributed to the classification performance with variable importance values ranging from 0 to ~14 (Table 4.9). Not surprisingly, the relative order of variable importance in the RF classifier had strong concordance with the measures of significance determined above.

4.3.2.2. ATC versus DTC array

The contingency table statistics identified 5 markers with significant associations between marker staining and pathology status for grouping 1 (Table 4.11) and 5 markers for grouping 2 (Table 4.12). It is interesting to note that both score groupings produced 5 significant markers, and though they do not perfectly overlap, all significant markers from both groupings were found to be significant by the MH test which does not depend on arbitrary score groupings (Table 4.11). This suggests that all eight significant markers from the MH test may be of biological relevance. In total, there were three up-regulated (over-expressed) and five down-regulated (under-expressed) markers in the ATC samples compared to the DTC samples. The clustering analysis and heat maps (Figure 4.6 and Figure 4.7) show that simple hierarchical clustering can separate ATC (light green side-bar) from DTC (dark green side-bar) with high accuracy, especially when the non-significant markers are excluded. The heat maps also provide a useful visualization of the marker staining patterns and show the three up-regulated (yellow side-bar) and five-down regulated (orange side-bar) markers clustered together as expected. The classification analysis identified the same eight markers (TG, MIB-1, CTNNB1, Bcl-2, E-CAD,

p53, VEGF, and TOPO-II) as the most important variables for discriminating patient-matched ATC from DTC samples with variable importance values ranging from 5 to 17 (Table 4.11). A classifier based on all markers was able to correctly classify ATC versus DTC with high accuracy, sensitivity and specificity. Specifically, cross-validation analysis reported an ROC integral of 0.983, an overall accuracy of 95.8%, sensitivity of 100% and specificity of 91.7%. Only a single DTC sample was misclassified as ATC.

4.4. Discussion

4.4.1. Meta-analysis

A common criticism of expression profiling studies is a lack of agreement between studies. However, by applying our meta-analysis method to a large number of published studies, we observe that many genes are consistently reported at a highly significant rate. These genes may represent real biologic effects that through repeated efforts have overcome the issues of noise and error typically associated with such experiments. A comparison of our meta-review method (using published gene lists) to a meta-analysis of a smaller subset of studies (for which raw data were available) showed a strong level of concordance. Thus, we believe our approach represents a useful alternative for identifying consistent gene expression markers when raw data are unavailable (as is generally the case). However, a limitation of our method resulting from unavailability of raw data is that we are unable to assign a measure of confidence at the gene level. We can identify consistently reported genes and rank them according to simple criteria like total sample size and average fold change, but we can not calculate a true combined fold-change or p-value. In order for more powerful meta-analysis methods to be applied, researchers must provide access to their raw data in the public databases and adhere to experimental annotation standards such as MIAME. This would also allow the thyroid cancer community to benefit from powerful toolkits such as the Oncomine resource which would essentially handle all the details of meta-analysis, and allow the researcher to focus on the problem of identifying clinically relevant diagnostic biomarkers. Also, we remind the reader that while we have focused on the 'cancer vs. non-cancer' comparisons in our discussion, a large number of other comparison groups were analyzed (Table 4.6) and we encourage the reader to explore these results on our supplementary webpage (http://www.bcgsc.ca/platform/bioinfo/ge/thyroid/).

As a means of further assessing our results, we discuss the significant gene ontology terms in terms of the published literature. We also review the top 12 'cancer vs. non-cancer' candidates to

identify which markers have been previously confirmed as differentially expressed or having diagnostic/prognostic utility in thyroid cancer (Table 4.13). In total, 10 of 12 markers have been confirmed at the RNA level and six of these have been validated at the protein level. For discussion purposes we have broken the genes into two categories, 'well-characterized' and 'novel or uncharacterized'. We also compare our results to a previous review of promising thyroid biomarkers and several subsequent studies.

4.4.1.1. Gene ontology analysis

The GO analysis of genes implicated by multiple cancer versus non-cancer studies provides a high-quality summary of biological themes that may be important in thyroid cancer (Table 4.8). Here we discuss the results for each GO term found to be significantly over-represented in our highest ranked genes. The binding of trace elements cadmium (Cd), copper (Cu) and selenium (Se) are likely important both individually and in combination in thyroid and other cancers. Cd is a known carcinogen that inhibits DNA repair [39] and has been implicated in numerous cancers including papillary thyroid carcinoma [40-42]. Elevated Cu levels have been observed in a number of neoplasms [43] and implicated in tumour angiogenesis [44] as well as oxidative DNA damage [45]. Se on the other hand is thought to have a cancer preventative action [46]. In animal models, retinoids have been shown to prevent or delay tumour promotion in a variety of cancers [47]. Retinoic acid (RA) has already been used successfully for redifferentiation therapy of thyroid cancer [48-50]. The mitogen-activated protein kinase pathway (MAPK) is known to act downstream of both the ras and ret genes, two genes with accepted links to the development of thyroid cancer. Recent studies have shown increased expression of activated, phosphorylated MAPK in papillary thyroid cancer; furthermore, they showed that pharmacological inhibition of the MAPK pathway *in vitro* reduced cellular proliferation in human thyroid cancer cell lines [51]. The role of extracellular matrix proteins in thyroid cancer has been studied to a significant extent. Tumour invasion requires degradation of extracellular matrix proteins by matrix metalloproteinases, thus there has been great interest in those proteins that might inhibit metalloproteinases. One such metalloproteinase inhibitor, TIMP-1 (identified in the meta-analysis), has been the focus of several studies and has been found to correlate with poorer prognosis in papillary thyroid cancer [52]. A large body of often-contradictory literature exists discussing the role of thyroid hormone levels in human disease. Thyroid hormone is an important regulator of growth, development, and differentiation and is mediated by thyroid receptors (TR) that act as transcription factors. In a recent review, Gonzalez-Sancho (2003)

outlines a number of studies that have implicated these genes in a number of cancers including thyroid cancer [53]. To summarize, the gene ontology terms associated with the meta-analysis genes represent processes and molecular functions consistent with the thyroid cancer literature.

4.4.1.2. Well-characterized biomarkers

We defined 'well-characterized' genes as those that have been validated in more than one follow-up study and at both the RNA and protein level such as MET, TFF3, SERPINA1, TIMP1, FN1, and TPO. Several studies have implicated MET protein expression in thyroid cancer as both a diagnostic marker [54-58] and prognostic marker [54, 56-58]. Increased MET expression has been associated with higher risk for metastasis [56] and recurrence in PTC [56, 57] and negative prognosis in FTC [58]. However, in another study, decreased MET was shown to be an effective predictor of distant metastases among PTC cases [54]. While no reports have evaluated TFF3 at the protein level, numerous studies have suggested TFF3 as a useful biomarker at the RNA level [25, 32, 59-62]. A two-gene panel of SFTPB and TFF3 was shown to correctly diagnose PTC with a sensitivity of 88.9%, specificity of 96.7% and accuracy of 94.9% [59]. TFF3/LGALS3 mRNA ratio was shown to distinguish FA from FTC with sensitivity and specificity of 80.0% and 91.5% respectively [62]. An antibody study of SERPINA1 reported immunoreactivity in 9/10 PTCs with no staining in the adjacent normal thyroid tissues [63]. TIMP1 up-regulation was confirmed by immunohistochemistry (IHC) with positive immunostaining in 68% of PTC cases and none of the normals [64]. Another IHC study of TIMP1 for 86 PTC specimens showed increased immunoreactivity in the tumour regions versus non-tumour regions in 92% of cases and significant correlations with unfavourable prognostic variables [52]. FN1 has been proposed as a useful RT-PCR marker of DTC [65] and an important modulator of thyroid cell adhesiveness and neoplastic cell growth [66]. An IHC study of 85 FTCs and 21 FAs reported that coexpression of FN1 and GAL3 or FN1 and HBME1 was restricted to cancer while their concurrent absence was highly specific (96%) for benign lesions [67]. A large number of studies have investigated TPO as a marker for thyroid carcinoma. Lazar *et al.* (2006) found that higher thyroid cancer stage was associated with lower TPO mRNA expression [68]. Segev *et al.* (2003) reviewed five IHC studies involving nearly 400 follicular lesions and found that 93% of FAs and 97% of FTCs were accurately diagnosed by TPO antibody staining. Studies using FNAB samples however have proved less promising with false positive rates as high as 32% [69]. For the most part, the six genes reviewed above appear

promising as thyroid cancer candidates and suggest our meta-analysis method is producing reasonable results.

4.4.1.3. Novel or uncharacterized biomarkers

For four genes (TGFA, QPCT, CRABP1 and FCGBP) we could find only a single follow-up study or validation experiment consistent with their potential importance in thyroid cancer. Bergstrom *et al*. (2000) suggest that increased expression of TGFA may be responsible for aberrant activation of EGFR and ultimately an overexpression and activation of MET (importance discussed above) [70]. Jarzab *et al.* (2005) built a classifier capable of discriminating between PTC and non-malignant samples in 90% of cases [27]. This classifier included QPCT (along with 18 other genes). QPCT was considered a novel gene and was validated by qPCR in that study but has been studied little further since. CRABP1 down-regulation was confirmed by RT-PCR (in one of the original microarray studies) [60] and another study reported that hypermethylation of promoter CpG islands for CRABP1 in PTC may explain the reduced expression [71]. Differential expression of FCGBP was confirmed in a separate study by restriction-mediated differential display and real-time RT-PCR [72].

For two genes (EPS8 and PROS1) we could find no confirmation beyond the initial microarray experiment. In our meta-analysis, five studies identified EPS8 [20, 26, 27, 32, 34] and four identified PROS1 [20, 26, 32, 73] as up-regulated in comparisons of cancer to non-cancer. And yet, to our knowledge, no follow-up study has confirmed either of these genes (even at the RNA level). It is unclear if genes like EPS8 and PROS1 have not been further validated because they are false-positives or simply because they have not yet been chosen for further study. These genes and the other less characterized candidates may represent novel diagnostic markers for thyroid cancer and warrant further investigation.

4.4.1.4. Comparison to previous and subsequent works

Comparison to a previous 'meta-review' by Segev *et al.* (2003) of mainly single-gene protein-level thyroid cancer studies found that four of their 12 markers identified as promising pre-operative diagnostic markers were identified as high-ranking candidates (top 30) in our meta-analysis (TPO, CD44, KRT19 and LGALS3) [69]. Two of their candidates were either not represented (HBME-1) or can not be reliably assayed by the microarray platforms (RET/PTC rearrangements). However, six other 'promising markers' (CDKN1B, TERT, CP/LTF,

DLGAP4, HMGA1, and PAX8) do have representation on at least some of the expression platforms and yet were not identified as differentially expressed in even a single study in our meta-analysis. These genes may have displayed some differential expression but not reached the required thresholds for inclusion in the published lists. Or, they may represent cases where changes in RNA levels do not correlate well with changes in protein levels. Segev *et al*. (2003) concluded that large scale thyroid tumour expression profiling of multiple markers on tumours from large and diverse patient cohorts are still required to identify a panel of markers with sufficient sensitivity and specificity to accurately diagnose indeterminate thyroid lesions [69]. We agree and believe that our meta-review of thyroid cancer gene expression profiling studies provides a high-quality list of candidates from which to identify such a panel. Furthermore, we have begun such a large scale profiling study of multiple markers on a large patient cohort using tissue microarrays (preliminary results discussed below).

Since our meta-analysis, three additional studies have been published that would likely have been included [74-76]. A brief look at these studies shows that the pattern of consistently reported genes has continued. Of the multi-study genes we identified as significant, 9 of the top 12 (reviewed above) and 18 of the top 39 (Table 4.7) genes were also reported in one or more of the recent studies. Two of the studies used an updated version of the Affymetrix HG-U95A platform (26 problem probe sets were omitted from HG-U95Av2 compared to HG-U95A) but the third used an entirely new platform (Applied Biosystems Human Genome Survey Microarray). Also subsequent to our meta-analysis, another meta-analysis was performed [77, 78]. In their study, Fujarewicz *et al.* (2007) [78] collected and analyzed a set of 180 microarray samples (90 HG-U95A and 90 HG-U133A) derived from 40 *de novo* microarray experiments, 124 previously published from their own lab [27, 79, 80], and 16 available publicly [32]. This set included 57 PTC, 61 benign thyroid tumour, and 62 normal thyroid tissue samples. They identified 43 genes which were most useful in the classification of malignant (PTC) from non-malignant (benign/normal). Their analysis was similar to our own subset meta-analysis for which we had raw Affymetrix data in terms of number of studies and sample sizes. However, only a small subset (the 16 publicly available microarrays [32]), overlapped between the two. On the other hand, 3 of the 4 previously published studies (representing 60/180 samples) were included in our overall meta-analysis from simple gene lists. Therefore, we expect some overlap between their gene list and ours. Indeed, we find a very high overlap with 12 of their genes found in our top 39 and an additional 21 at lower ranks in our meta-analysis. Of these, three of their highest

ranked genes (ranks 3, 4 and 6) were in our top 12 (MET, FN1, and QPCT). In fact, only 9 of 42 of their genes were absent entirely (not found in any study in our meta-analysis). It was also reassuring to observe that in all cases of overlap but one (32 of 42 genes) the two studies also agree on direction of change. The results above again demonstrate that completely different expression profiling technologies and analysis methods consistently identify a common set of differentially expressed genes for thyroid cancer.

It should be noted that additional genes might reach significant overlap in our meta-analysis with the inclusion of new datasets. As new datasets become available, it may be useful to update the meta-analysis. This would be greatly facilitated if resources such as Oncomine would open themselves to greater public input and collaboration. The current version (Oncomine 3.0) contains 289 studies representing 20,835 samples and 39 cancer types and allows meta-analysis of differential expression for any set of studies that the user selects [81]. This is a very powerful resource that removes many of the practical barriers (discussed above), and makes the final product (the significant genes with multi-study overlap) directly available to the researcher community. Unfortunately, the database does not follow an open model whereby users can submit their own datasets directly, access the database programmatically, or download raw data. Currently (June, 2007), only a single thyroid cancer study has been added to this system. This is likely, at least in part, because many of the groups reporting thyroid cancer profiling studies have not submitted their raw data to the public databases or responded to requests. As mentioned above, for our seven email requests we received only two favourable responses.

4.4.2. Tissue microarray

4.4.2.1. Malignant versus benign

By comparing a large cohort of benign lesions to malignant tumours we identified a large number of significantly altered markers. Even after multiple testing corrections, 33 of the 56 markers tested showed some difference in expression. Because of greater sample numbers much weaker effects were able to reach statistical significance on the malignant/benign array compared to the ATC/DTC array. While all significant alterations are of potential interest, we will limit the discussion here to just those with a variable importance of 5 or greater in the random forests classifier (similar to that observed for the eight significant ATC markers discussed below). We will also discuss the classifier performance in relation to the current literature.

Of the five markers found to be most significantly altered, VEGFA (VEGF) was down-regulated and LGALS3 (Galectin-3), KRT19 (CK19), AR, and AURKA (Aurora-A) were up-regulated. VEGFA is a multifunctional cytokine with secretion that is regulated by a variety of cytokines and growth factors, plays an important role in angiogenesis, and is over-expressed in many human malignancies [82]. Angiogenesis is important for both local tumour growth and distant cancer spread [83]. Huang *et al.* (2001) evaluated VEGFA expression in a cohort of 117 thyroid tumours, which included 76 PTC, 12 FTC, 13 FA, 6 HCC, 2 HCA, and 8 ATC samples [84]. They found VEGFA levels were usually higher in follicular adenoma (FA) than in follicular carcinoma (FTC), consistent with our observation of down regulation in malignant samples compared to benign. However, they also report consistently strong and diffuse staining of VEGFA in all PTCs whereas we saw positive staining in only 37.5% of malignant (predominantly PTC), significantly less than the 90% positive staining for benign. Other studies have shown that VEGFA expression was higher in cancer cell lines or primary tumours than normal thyroid and higher in metastatic tumours than non-metastatic tumours [85, 86]. Most recently, a case-control study (332 cases, 261 controls) identified a regulatory SNP (in the promoter region and previously shown to be related to VEGF expression) which was associated with increased risk of thyroid cancer and lymph node metastasis in men (OR=1.97, 95% CI 1.16-3.37, p=0.013).

Of all the markers that have been evaluated for diagnosis of thyroid cancer, LGALS3 (Galectin-3) has been the most widely studied. Galectins are involved in many of the biologic functions of the cell including: growth, differentiation, adhesion, mRNA processing, and apoptosis [87]. Multiple investigators have focused on LGALS3, a chimera type galectin that contains a nonlectin portion connected to a lectin domain, as a diagnostic marker for thyroid cancer because it is consistently expressed in the cytoplasm of malignant thyrocytes [88]. In a recent multicenter study Bartolazzi *et al*. (2001) reported galectin-3 expression in 1,009 thyroid specimens to be a sensitive and specific marker for thyroid cancer diagnosis [89]. KRT19 (cytokeratin-19) is a cytoskeletal protein that is highly expressed in PTC [90] and has been reported as useful in distinguishing this from benign or other kinds of malignant thyroid tissue in several studies [91-93]. The up-regulation of LGALS3 and KRT19 in thyroid carcinoma reported in the literature is consistent with what we observed in both the meta-analysis and later in the TMA analysis (see further discussion in section 4.4.2.3). Both have also since shown promise in a number of diagnostic panels under development (see discussion below).

Very few studies have investigated expression of AR (Androgen receptor) or AURKA (Aurora-A) in thyroid cancer to date. Androgen receptor functions as a steroid-hormone activated transcription factor and plays a central role in prostate cancer progression [94]. A single small study (n=28) has suggested that medullary thyroid carcinoma (MTC) might be influenced by sex steroid hormones through the expression of ERβ and AR on C cells [95]. Our findings may be the first to suggest a more general role for AR in differentiated thyroid carcinoma. The aurora kinases (AURKA, AURKB, AURKC) are involved in the regulation of cell cycle progression and alterations in their expression have been associated with malignant transformation [96, 97]. The first study to consider aurora kinase expression showed that all three were up-regulated (compared to normal thyrocytes) in human cell lines derived from malignant thyroid carcinoma (FTC, PTC and ATC) but not benign (FA). Thus, the TMA analysis presented in this chapter is the first to confirm up-regulation in primary tissues of an aurora kinase in malignant differentiated thyroid cancer. In a TMA study of 32 ATC cases, Wiseman *et al.* (2007) showed expression of AURKA in 83% of cases [98]. However, in the ATC/DTC comparison described in this chapter, no change in expression for Aurora-A was observed.

To summarize, we investigated the expression of 56 proteins in 100 benign and 96 malignant thyroid lesions by tissue microarray analysis. A number of these (33) showed significant alterations in their expression. Among the top five, three were very well-characterized thyroid cancer biomarkers (VEGF, LGALS3, and KRT19) and two were relatively novel (AR and AURKA). Based on a review of the literature, our TMA study appears to represent the largest number of antibodies (>50) investigated by TMA for a large patient cohort (~200 patients). It produced a classifier with performance comparable to others in the field as detailed below.

In addition to investigating promising markers individually, many groups have attempted to improve the sensitivity and specificity of thyroid cancer diagnosis by evaluating panels of molecular markers. Some have identified large multi-gene panels by gene expression profiling studies whereas others have focused on smaller panels with more targeted assays. In the largest multi-gene microarray approach, Fujarewicz *et al.* (2007) compiled a collection of 180 microarray experiments (57 PTC, 61 benign and 62 normal), from multiple labs and developed a 20-gene classifier with an accuracy of 98.5% for PTC diagnosis. Their most useful markers included well-known PTC markers such as MET, FN1, DPP4 and ADORA1 as well as a number

of novel ones and had large overlap with our own meta-analysis (as discussed in section 4.4.1.4). A six-gene diagnostic panel (kit, Hs.296031, Hs.24183, LSM7, SYNGR2, and C21or4) was demonstrated to have classification potential in a microarray study and then confirmed by qRT-PCR as being able to differentiate between benign and malignant thyroid tumours with high sensitivity and specificity (correctly predicting 9/10 unknowns) [99]. Evaluation of 85 carcinomas (67 PTC, 6 FTC, 8 HCC, and 4 ATC) and 21 adenomas by IHC found that coexpression of more than one of FN1, LGALS3 and HBME1 was seen in 95% of carcinomas whereas their concurrent absence was highly specific (96%) for benign lesions [67]. Another study reported coexpression of HBME1 and LGALS3 in 36 of 42 PTCs and none of the 58 benign samples assayed by IHC [100]. Some groups have even advanced to testing their panels on FNAB samples instead of post-operative resections, an important step for migrating a test to the clinic. Kebebew *et al.* (2006) showed that an assay using real-time qRT-PCR of ECM1, TMPRSS4, ANGPT2, and TIMP1 could achieve sensitivity of 91.0% and specificity of 95.0% for diagnosis of 31 thyroid nodule FNAB samples [101]. Lubitz *et al.* (2006) recently used DNA microarray analysis to identify 25 differentially expressed genes in a comparison of 26 benign and 24 malignant thyroid carcinomas [76]. Unsupervised hierarchical clustering was used to classify 22 FNAB specimens. The classification was 100% concordant to the final histological diagnosis compared to 76% from preoperative cytological FNAB diagnosis. Other promising panels of diagnostic markers have included: DDIT3, ARG2, C1orf24, ITM1, LGALS3, HBME-1, EPHB2 (ERK), RET, CDKN2A (p16), KRT19, NKX2-1 (TTF-1), CITED1, and S100A4 to name a few [92, 93, 102-105]. Thus, we can see that several markers identified in either our meta-analysis or tissue microarray analyses are already making progress. Perhaps some of these panels or a combination of them should soon be advanced to larger trials.

4.4.2.2. ATC versus DTC

Anaplastic transformation represents a post-malignant tumour progression. Specifically, transformation is a terminal event, with ATC representing the end point of thyroid tumour evolution [14]. In the current study, by comparing ATCs with the DTCs from which they arose, we have identified 8 significantly altered markers (3 up-regulated and 5 down-regulated) potentially involved in the transformation process.

The 3 markers identified as being significantly up-regulated, or over-expressed, in ATC compared to DTC were TP53 (p53), MIB1 (MIB-1), and TOP2A (topoisomerase II-α). TP53 is a

tumour suppressor gene often referred to as the "guardian of the genome," and approximately half of human tumours exhibit a TP53 alteration [106, 107]. Its gene product, the p53 protein, acts as a transcription factor that regulates downstream genes that are involved in DNA repair, cell cycle arrest, and apoptosis [106, 107]. The status of TP53 has also been linked to tumour chemosensitivity, radiosensitivity, and prognosis of many human cancer types [106-109]. TP53 is also believed to play an important role in thyroid cancer transformation because it is rarely altered in DTC but commonly mutated in ATC. In a multi-study review of 265 ATC cases, 52% of tumours exhibited either gene or protein-level alteration of TP53 [110]. The important role of TP53 in anaplastic transformation has led several investigators to report the development of DTC characteristics with reintroduction of wild-type p53 into ATC [14]. MIB1 antibody binds to the Ki-67 nuclear antigen, and its level of expression correlates with measurements of cellular proliferation [111]. Kjellman *et al.* (2003) evaluated MIB1 expression with IHC analysis in a cohort of 144 thyroid tumours (including 40 DTCs and 8 ATCs) and found that expression was higher in ATCs (median, 16.2%) when compared with DTCs (median, 1.9% in PTCs and 2.7% in FTCs) [112]. TOP2A is a nuclear enzyme that is required for chromatin condensation and segregation during mitosis and is expressed in the S, G2, and M phases of the cell cycle [113, 114]. As a consequence of the expression of TOP2A being coupled with the cell cycle, like MIB-1, it is also considered a marker of cellular proliferation [113, 114]. Notably, TOP2A may serve as a target for many anticancer drugs, including anthracyclines, and the presence of gene amplification, deletion, or protein expression may predict response to treatment [115, 116]. Currently, the topoisomerase II inhibitor doxorubicin, which has a reported response rate of 5% to 22%, is considered to be the most effective drug for treatment of ATC [117-119]. Yet, few reports have evaluated TOP2A expression by thyroid tumours. Using immunohistochemical analysis, Lee *et al.* (2000) observed a higher TOP2A expression in ATC compared with DTC [114]. Fluge *et al.* (2006) identified TOP2A to be markedly up-regulated in clinically aggressive, poorly differentiated DTCs when compared with typical DTCs [120]. Poorly differentiated DTC is considered an intermediate form in the progression of DTC to ATC [121].

The 5 markers identified as being significantly down-regulated, or showing decreased expression in ATCs compared with the DTCs from which they transformed, were TG (thyroglobulin), CDH1 (E-cadherin), CTNNB1 (β-catenin), BCL2 (Bcl-2), and VEGFA (VEGF). The expression of TG is thought to be suggestive of a differentiated thyroid tumour phenotype, and TG protein expression is commonly absent in ATC [122]. CDH1 is a transmembrane glycoprotein that is

important for cell-to-cell adhesion and complexes with catenin proteins via an intracellular domain. Either CTNNB1 or CTNNG (γ-catenin) binds to a common CDH1 domain and then binds to CTNNA1 (α-catenin), which anchors the actin cytoskeleton to the cadherin-mediated adhesion complex [123]. CTNNB1 also functions as a regulator of cell growth and survival as a downstream effector of the Wnt signalling pathway [124]. Wiseman *et al.* (2006) previously evaluated and reviewed evidence that suggested that derangement of the E-cadherin–catenin complex is involved in anaplastic transformation of thyroid cancer [125]. The protein encoded by the BCL2 proto-oncogene is responsible for prolongation of cell survival by blocking apoptosis [126-128]. In a study that evaluated 134 thyroid tumours for Bcl-2 protein expression with IHC analysis, Pollina *et al*. (1996) reported down-regulation of BCL2 in ATC [126]. In this study, BCL2 immunoreactivity was identified in 60 of 70 DTCs (85.7%) and only 8 of 24 ATCs (33.3%). Similar observations have been reported by other investigators [127]. In an ATC cell line model, Kim *et al.* (2003) reported success in using a Bcl-2 antisense oligonucleotide to enhance apoptosis and increase ATC drug sensitivity [128]. The general involvement of VEGFA in human malignancy and thyroid cancer was discussed above (section 4.4.2.1). Referring again to the study by Huang *et al.* (2001) [84] of VEGFA expression, they found that all PTCs exhibited strong diffuse staining, whereas the ATCs showed weak and infrequent immunoreactivity. In another study that examined 52 DTCs (34 PTCs and 18 FTCs) and 8 poorly differentiated DTCs, Vieira at al reported VEGFA expression to be significantly more prevalent in PTC (79%) than either FTC (50%) or poorly differentiated DTC (37%) [129]. Although the underlying mechanisms are unknown, these observations suggest a potentially important role for alternate pathways of angiogenesis in these tumours. Currently, anti-angiogenesis agents, including drugs that target VEGF and other mediators of angiogenesis, are being studied for the treatment of ATC [130-132].

To summarize, the observed up-regulation of TP53, MIB1, and TOP2A and down-regulation of TG, CDH1, CTNNB1, BCL2, and VEGFA in the progression of DTC into ATC is consistent with current literature reports and further highlights the importance of these genes in thyroid tumour progression. By using a TMA-based approach and evaluating 12 coexisting adjacent DTC and ATC tumours, we have evaluated the change in expression profile for a panel of 63 molecular markers, with the aim of identifying the molecular alterations that occur during the transformation of DTC into ATC. Not only did analysis of the tumour expression profiles reveal 8 markers as being significantly altered when comparing patient-matched ATC and DTC

samples, but independent of tumour histologic features, a classifier based on all markers was able to correctly differentiate ATC and DTC with high accuracy, sensitivity, and specificity. While distinguishing undifferentiated from differentiated tumour tissue is not a major challenge for the clinician, the classification approach represents a useful method for identifying the discriminating (and therefore biologically interesting) markers. The specific intratumoural molecular alterations that occur during transformation warrant further evaluation as molecular prognosticators for DTC. The DTC patient risk stratification systems currently guide the extent of surgery and the use of adjuvant radioactive iodine therapy, although no single system has been universally accepted or applied [133]. Further evaluation of the prognostic utility of one or more of the 8 markers we have identified as being significantly altered during transformation, in large DTC patient cohorts, could potentially lead to improved treatment selection and patient outcomes. The molecular markers we and others have identified as being significantly altered in transformation may also represent important targets for the treatment of ATC. Molecular targets that are important for the transformation of DTC could potentially prevent ATC development or, as has been demonstrated in the laboratory with reintroduction of p53 into ATC cell lines, lead to the development of more differentiated tumour characteristics [14].

4.4.2.3. Comparison of malignant/benign and ATC/DTC tissue microarray results

In a sense, the differentiated thyroid carcinomas in the ATC versus DTC comparison are equivalent to the malignant samples in the malignant versus benign comparison. We could consider thyroid cancer as a progression from benign to differentiated malignant pathology and then (rarely) to undifferentiated anaplastic cancer. Therefore, it might be of interest to examine the overlap between altered markers for the two types of comparisons. Of the eight significant markers in the ATC versus DTC samples, seven were also assayed on the benign versus malignant array. MIB-1 was not yet assayed on the benign/malignant array at time of writing. Of the seven overlapping markers, five showed a significant alteration in both comparisons. TP53 and TOPO2A, two of the up-regulated markers in ATC, showed no difference in expression between benign and malignant. In fact both were barely expressed in the 196 benign/malignant patients at 5.1% and 1.0% compared to 83.3% and 91.7% expression in ATC (percentages reflect score grouping 1). This could indicate activation of processes or pathways involved specifically in transformation to ATC but not in the initial development of malignancy. All five down-regulated genes in ATC also showed some alteration in malignant patients compared to benign. However, for TG and ECAD the effects were very weak in comparison to the difference seen in

ATC. Both genes actually show very high expression in benign/malignant samples with overall positive staining of 100% and 99.0% respectively for score grouping 1 and only moderate down-regulation for grouping 2 and ungrouped statistics. But, a sharp drop in expression of these markers was observed from DTC to ATC (91.7% to 9.1% and 91.7% to 16.7% respectively). These could represent genes where weak down-regulation in DTC (or down-regulation in a patient subset) foreshadows future transformation to more aggressive disease. CTNNB1 on the other hand shows the opposite effect between the two comparisons. In malignant samples there was a slight up-regulation of CTNNB1 compared to benign whereas in ATC we see a down-regulation compared to DTC. However, the difference in both comparisons is moderate making it hard to draw confident conclusions from this observation. Most interesting perhaps are BCL2 and VEGF which showed relatively strong down-regulation alterations in both types of comparisons. It is conceivable that decreased expression of these genes is linked to both progressions: from benign to malignant/DTC and then from DTC to ATC.

4.4.2.4. Performance of meta-analysis markers on tissue microarray

A number of markers identified in the meta-analysis had antibodies available on hand for use on the tissue microarray. These covered a wide range of meta-analysis ranks and therefore give us the opportunity to generally evaluate the utility of the meta-analysis for predicting useful markers in tissue microarray analysis. Antibodies for a selection of the top 12 meta-analysis markers have also been ordered and are being processed but were not ready at time of writing. The performance of meta-analysis markers (cancer versus non-cancer comparison group) on the malignant versus benign TMA are summarized in Table 4.14. All genes, even those which were identified in only a single expression profiling study (with a very low rank in our meta-analysis) are included. First, it is interesting to note that every gene which was identified in the meta-analysis (even those with lower ranks) showed a significant alteration on the TMA. This indicates that overall, differential expression at the RNA level is a fairly good predictor of differential expression at the protein level. However, the relatively low correlation between meta-analysis rank and TMA rank (Spearman's correlation; r=0.164) indicates that there is a wide range in performance. For the four genes (SERPINA1, KRT19, LGALS3, and CCND1) which we considered most reliable from the meta-analysis (support from 3 or more studies), three were in the top 10 in terms of TMA performance. Two of these (LGALS3 and KRT19) were by far the most useful TMA markers in terms of classifier performance. Conversely, one the best meta-analysis markers (SERPINA1), reported in six independent RNA profiling studies,

had relatively poor performance on the TMA. It will be interesting to see how other 'top 12' meta-analysis markers perform. It is possible that this RNA-level change is only partially translated to a protein-level change or that this particular antibody does not work well. Another interesting observation from Table 4.14 is that the next most useful TMA marker (CDH1 with TMA rank 9) had the worst meta-analysis rank in the list. Looking more closely, it was determined that the CDH1 actually had an overlap of 2 and would have ranked at 85 instead of 746, but was heavily penalized because the two studies reporting it as differentially expressed did not agree on the direction of change (i.e., one listed it as up- and the other as down-regulated). Looking at the specific studies in this case, one was comparing papillary thyroid carcinoma (PTC) to normal tissue [30] whereas the other was comparing a mixture of PTC and follicular variant PTC (FVPTC) to a mixture of benign lesions (FA and HN) [134]. Thus, it is easy to imagine that CDH1 might truly be up-regulated in one of these comparisons and down-regulated in the other. This 'spurious discrepancy' is a consequence of grouping studies into the overly general 'cancer versus non-cancer' analysis. Fixing the rank for CDH1 improves the TMA/meta-analysis correlation quite significantly to 0.318. In the future, it might be advisable to remove the penalty on 'direction disagreements' for such cases. Indeed, where multiple studies report a gene as differentially expressed, even if they disagree on the direction, this gene might still be a better candidate for further validation than a gene identified only once. In any case, it seems clear that genes predicted by gene expression profiling studies, and in particular those predicted by multiple studies, are a rich source of candidates for further investigation by tissue microarray. It is our hope that additional markers currently being processed will aid in the development of a clinically useful classifier for distinguishing benign from malignant thyroid lesions.

4.5. Conclusions

Thyroid cancer remains a disease with serious challenges in terms of both diagnostic determination and prognostic outcome (especially in cases of transformation into the rare but deadly anaplastic thyroid carcinoma). Molecular profiling studies have identified a plethora of differentially expressed genes with potential to address these challenges. Our meta-analysis of these studies has shown that a significant number of genes are consistently reported by different platforms and independent research groups. Some of these "multi-study" genes have since been verified as having diagnostic utility, whereas others have received relatively little attention and represent potentially novel biomarkers. The meta-analysis method itself represents a novel and

useful approach for identifying potential biomarkers from multiple expression profiling studies when the raw data are unavailable. It has been referenced several times in the thyroid cancer literature and has already been applied to other diseases such as colon cancer [144].

While a number of promising biomarkers or biomarker panels have been identified, a simple, consistent and accurate molecular signature for the preoperative diagnosis of thyroid malignancy has remained elusive. Large scale thyroid tumour studies, evaluating many markers on lesions from large and diverse patient cohorts, are still required to identify a panel with sufficient sensitivity and specificity to accurately diagnose individuals with malignancy and the subset that may progress to more aggressive disease. Tissue microarrays represent a powerful method to achieve this aim. Few tissue microarray studies of cancer have investigated large numbers of antibodies or applied microarray analysis methods such as clustering and classification to them. Here, we have presented preliminary tissue microarray results on two patient cohorts comparing malignant versus benign and undifferentiated (ATC) to differentiated (DTC) thyroid lesions for a large number of antibodies. In both cases we identified alterations in a broad range of important underlying cellular processes and created classifiers with high accuracy, sensitivity and specificity. Based on the performance of antibodies which we had on hand, inclusion of additional high-ranking markers from the meta-analysis should further improve the classifier performance and may lead to a diagnostic panel of clinical utility.

Table 4.1. Thyroid cancer profiling studies included in meta-analysis

The 'genes/features' column refers to the number of clones or probes spotted on the various cDNA or oligonucleotide arrays. SAGE is listed as not applicable (N/A) because the number of unique tags observed is indeterminate and depends on the library sequencing depth. The number of samples assayed for each condition in each comparison is shown in brackets in the 'condition 1/2' columns. In cases where multiple comparisons were conducted for a single study, a sample can be listed more than once. The numbers of 'up-/down-regulated' genes reported are for condition 1 relative to condition 2 for each comparison as provided by the publication, supplementary materials, or personal communications with the authors (except for Giordano *et al.* (2005) [20]; see methods). The numbers of up- or down-regulated genes after mapping to Entrez gene ID are listed in brackets. Only genes that could be mapped to a common identifier were used in our subsequent overlap analyses (see Methods).

Study	**Platform**	**Genes/ features**	**Comparison**		**Up-/down-regulated features (mapped genes)**
			Condition 1 (No. samples)	**Condition 2 (No. samples)**	
Arnaldi *et al*. 2005[28]	Custom cDNA array	1807	FCL(1)	Norm (1)	9/20 (9/17)
			FCL(1), PCL(1), UCL(1)	Norm (1)	3/6 (3/3)
			PCL(1)	Norm (1)	1/8 (1/8)
			UCL(1)	Norm (1)	1/7 (1/6)
Giordano *et al*. 2005[20]	Affymetrix HG-U133A	22283	PTC(51)	Norm(4)	90/151 (69/122)
Jarzab *et al*. 2005[27]	Affymetrix HG-U133A	22283	PTC(16)	Norm(16)	75/27 (71/26)
Weber *et al*. 2005[34]	Affymetrix HG-U133A	22283	FA(12)	FTC(12)	12/84 (12/65)
Aldred *et al*. 2004[33]	Affymetrix HG-U95A	12558	FTC (9)	PTC(6), Norm(13)	0/142 (0/126)
			PTC (6)	FTC(9), Norm(13)	68/0 (59/0)
Cerutti *et al*. 2004[135]	SAGE	N/A	FA(1)	FTC(1), Norm(1)	5/0 (4/0)
			FTC(1)	FA(1), Norm(1)	12/0 (9/0)
Chevillard *et al*. 2004[29]	custom cDNA array	5760	FTC(3)	FA(4)	12/31 (12/30)
			FVPTC(3)	PTC(2)	123/16 (123/16)
Finley *et al*. 2004[25]	Affymetrix HG-U95A	12558	PTC(7), FVPTC(7)	FA(14), HN(7)	48/85 (48/82)
Finley *et al*. 2004[26]	Affymetrix HG-U95A	12558	FTC(9), PTC(11), FVPTC(13)	FA(16), HN(10)	50/55 (49/52)
Hawthorne *et al*. 2004[60]	Affymetrix HG-U95A	12558	GT(6)	Norm(6)	1/7 (0/6)
			PTC(8)	GT(6)	10/28 (8/18)
			PTC(8)	Norm(8)	4/4 (3/3)
Mazzanti *et al*. 2004[134]	Hs-UniGem2 human cDNA array	9984	PTC(17), FVPTC(15)	FA(16), HN(15)	5/41 (4/35)
Onda *et al*. 2004[18]	Amersham custom cDNA array	27648	ACL(11), ATC(10)	Norm(10)	31/56 (27/54)
Pauws *et al*. 2004[136]	SAGE	N/A	FVPTC(1)	Norm(1)	33/9 (14/4)
Yano *et al*. 2004[30]	Amersham custom cDNA array	3968	PTC(7)	Norm(7)	54/0 (41/0)
Zou *et al*. 2004[137]	Atlas human cancer cDNA array (cancer 1.2 array)	1176	MACL(1)	ACL(1)	43/21 (42/20)
Barden *et al*. 2003[73]	Affymetrix HG-U95A	12558	FTC(9)	FA(10)	59/45 (53/42)
Wasenius *et al*. 2003[64]	Atlas human cancer cDNA array (cancer 1.2 array)	1176	PTC(18)	Norm(3)	12/9 (12/8)
Chen *et al*. 2001[138]	Atlas human cDNA array (Clontech)	588	M (1)	FTC (1)	18/40 (17/40)
Eszlinger *et al*. 2001[80]	Atlas human cDNA array (Clontech)	588	AFTN(3), CTN(3)	Norm(6)	0/16 (0/12)
Huang *et al*. 2001[32]	Affymetrix HG-U95A	12558	PTC (8)	Norm (8)	24/27 (24/27)
Takano *et al*. 2000[19]	SAGE	N/A	FTC(1)	ATC(1)	3/10 (1/7)
			FTC(1)	FA(1)	4/1 (2/1)
			Norm(1)	FA(1)	6/0 (2/0)
			PTC(1)	ATC(1)	2/11 (0/8)
			PTC(1)	FA(1)	7/0 (2/0)
			PTC(1)	FTC(1)	2/1 (1/1)
21 studies	**10 platforms**		**34 comparisons (473 samples)**		**827/958 (723/839)**

Table 4.2. List of Abbreviations for thyroid samples

Lists of abbreviations used for thyroid samples. Previously reported conventions were followed wherever possible.

ACL	Anaplastic thyroid cancer cell line
AFTN	Autonomously functioning thyroid nodules
ATC	Anaplastic thyroid cancer
CTN	Cold thyroid nodule
DTC	Differentiated thyroid cancer
FA	Follicular adenoma
FCL	Follicular carcinoma cell line
FTC	Follicular thyroid carcinoma
FVPTC	Follicular variant papillary carcinoma
GT	Goiter
HCC	Hurthle cell carcinoma
HCA	Hurthle cell adenoma
HN	Hyperplastic nodule
HT	Hashimoto's Thyroiditis
LT	Lymphocytic Thyroiditis
M	Metastatic
MACL	Anaplastic thyroid cancer cell line with metastatic capacity
MTC	Medullary thyroid carcinoma
Norm	Normal
PCL	Papillary carcinoma cell line
PTC	Papillary thyroid carcinoma
TCVPTC	Tall-cell variant PTC
UCL	Undifferentiated carcinoma cell line

Table 4.3. Antibodies and Scoring Systems Used for TMA analysis

The antibody name is used in all tables and figures relating to the TMA analysis. The target gene refers to the gene locus thought to be targeted by each antibody and utilizes Entrez Gene's official symbol as identifier wherever possible. Scoring types are explained in Table 4.5. Abbreviations: c, cytoplasmic; m, membranous; n, nuclear. For the arrays probed, A is for the Anaplastic array (matched ATC versus DTC) and B is for the malignant versus benign array. As indicated, most but not all antibodies were tested on both arrays.

Antibody name	Target Gene Symbol	Target Gene Name	Localization	Scoring Type	Array (s) probed
AAT	SERPINA1	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	c	F	AB
AMF-R	AMFR	Autocrine motility factor receptor	c	B	AB
AR	AR	Androgen receptor	n	G	AB
Aurora-A	AURKA	Aurora kinase A	c	E	AB
Aurora-B	AURKB	Aurora kinase B	n	B	A
Aurora-C	AURKC	Aurora kinase C	c	C	AB
Bcl-2	BCL2	B-cell CLL/lymphoma 2	c	C	AB
CAIX	CA9	Carbonic anhydrase IX	c ± m	G	A
CAV-1	CAV1	Caveolin 1, caveolae protein, 22kDa	m	I	AB
Caveolin	CAV1	Caveolin 1, caveolae protein, 22kDa	m	I	AB
CDX2	CDX2	Caudal type homeobox 2	n	B	AB
CK19	KRT19	Keratin 19	c	G	AB
c-kit	KIT	V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	c ± m	B	AB
Clusterin	CLU	Clusterin	c ± m	B	AB
COX2	PTGS2	Prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	c	K	AB
CR3	ITGAM	Integrin, alpha M (complement component 3 receptor 3 subunit)	n	H	AB
CTNNB1	CTNNB1	Catenin (cadherin-associated protein), beta 1	c ± m	G	AB
Cyclin-D1	CCND1	Cyclin D1	n	F	AB
Cyclin-E	CCNE1	Cyclin E1	n	F	AB
E-CAD	CDH1	Cadherin 1, type 1, E-cadherin (epithelial)	m	C	AB
EGFR	EGFR	Epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	c ± m	G	AB
ER	ESR1	Estrogen receptor 1	n	C	AB
Galectin-3	LGALS3	Galectin-3	c	G	AB
HBME-1	N/A	N/A	c ± m	D	AB
HER2	ERBB2	V-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	m	A	AB
HER3	ERBB3	V-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)	c ± m	G	AB
HER4	ERBB4	V-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)	c ± m	C	AB
HSP-27	HSPB1	Heat shock 27kDa protein 1	c	H	AB
IGF1-R	IGF1R	Insulin-like growth factor receptor 1	m	B	A
IGFBP2	IGFBP2	Insulin-like growth factor binding protein 2	c	I	B
IGFBP5	IGFBP5	Insulin-like growth factor binding protein 5	c	I	B
ILK	ILK	Integrin-linked kinase	m	C	A
INH	INHBB	Inhibin	n	F	AB
KI67	MKI67	Antigen identified by monoclonal antibody Ki-67	n	F	B
MDM2	MDM2	Mdm2, transformed 3T3 cell double minute 2, p53 binding protein (mouse)	c	D	AB
MIB-1	MIB1	Mindbomb homolog 1 (Drosophila)	n	F	A
MLH1	MLH1	MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	n	J	AB
MSH2	MSH2	MutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)	n	J	A
MSH6	MSH6	MutS homolog 6 (E. coli)	n	J	A
O13	CD99	Antigen identified by monoclonal antibodies 12E7, F21 and O13	n	H	AB
P16	CDKN2A	Cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	n	E	AB
P21	CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	n	F	AB

Antibody name	Target Gene Symbol	Target Gene Name	Localization	Scoring Type	Array (s) probed
P27	CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)	n	C	AB
P504S	AMACR	Alpha-methylacyl-CoA racemase	c	G	AB
P53	TP53	Tumour protein p53 (Li-Fraumeni syndrome)	n	E	AB
P57	CDKN1C	Cyclin-dependent kinase inhibitor 1C (p57, Kip2)	n	B	AB
P63	TP63	Tumour protein p63	n	H	AB
P75-NTR	NGFR	Nerve growth factor receptor (TNFR superfamily, member 16)	c ± m	G	AB
P-AKT	AKT	V-akt murine thymoma viral oncogene homolog 1	c ± m	B	A
PGI	GPI	Autocrine motility factor	c	C	AB
PMS2	PMS2	PMS2 postmeiotic segregation increased 2 (S. cerevisiae)	n	J	AB
PR	PGR	Progesterone receptor	n	c	AB
PSA	KLK3	Kallikrein-related peptidase 3	n	H	AB
RET	RET	Ret proto-oncogene	c	C	AB
S100	S100B	S100 calcium binding protein B	c	E	AB
Syntrophin	N/A	N/A	c	I	AB
TDT	DNTT	Deoxynucleotidyltransferase, terminal	n	H	AB
TG	TG	Thyroglobulin	c	D	AB
TOPO-II	TOP2A	Topoisomerase (DNA) II alpha 170kDa	n	G	AB
TS106	TYMS	Thymidylate synthase	n	G	AB
TSH	TSHB	Thyroid stimulating hormone, beta	c	B	AB
TTF-1	NKX2-1	NK2 homeobox 1	n	F	AB
UPA-R	PLAUR	Plasminogen activator, urokinase receptor	c	C	A
VEGF	VEGFA	Vascular endothelial growth factor A	c ± m	C	AB
WT1	WT1	Wilms tumour 1	n	F	AB

Table 4.4. Characteristics of antibodies used for TMA analysis

Antibody name	Isotype	Clone	Company	Catalog No.	Antigen Retrieval	Conc.
AAT	Rabbit anti-human	NA	Dako, Glostrup, Denmark	A0012	None	1:10000
AMF-R	Rat IgM	NA	R. Nabi, University of British Columbia, Vancouver, British Columbia, Canada	NA	Citrate buffer (pH, 6)	1:50
AR	Mouse IgGκ	F39.4.1	Biogenex Laboratories, San Ramon, California	MU256-UC	S30 EDTA	1:250
Aurora-A	Rabbit polyclonal	NA	Santa Cruz Biotechnology, Santa Cruz, California	SC-14318	Heat induced	1:20
Aurora-B	Mouse monoclonal	6	BD Transduction Laboratories, San Jose, California	611082	Heat induced	1:20
Aurora-C	Rabbit polyclonal	NA	Zymed Laboratories, San Francisco, California	38-9400	Heat induced	1:15
Bcl-2	Mouse IgG1κ	124	Dako	M0887	Heat induced	1:20
CAIX	Mouse monoclonal	M75	S. Chia, BC Cancer Agency, Vancouver	NA	Citrate buffer (pH, 6)	1:20
CAV-1	Rabbit polyclonal	NA	Santa Cruz Biotechnology	SC-894	Heat induced	1:1000
Caveolin	Rabbit polyclonal	NA	BD Transduction Laboratories	610059	Heat induced	1:200
CDX2	Mouse monoclonal IgGκ	AMT 28	Novocastra Laboratories, Norwell, Massachusetts	153505	Heat induced	1:50
CK19	IgGκ1	Ba17	Dako	MO772	PC8 citrate	1:100
c-kit	Rabbit polyclonal	NA	Dako	A4502	Heat induced	1:100
Clusterin	Goat polyclonal	M-18	Santa Cruz Biotechnology	SC-6240	Citrate buffer (pH, 6)	1:600
COX2	Rabbit IgG	SP21	Lab Vision Corp, Fremont, California	RM-9121-S	PC8 citrate	1:100
CR3	Mouse monoclonal	43	Dako	M0775	Heat induced	1:500
CTNNB1	Mouse monoclonal	14	BD Transduction Laboratories	610153	Citrate buffer (pH, 6)	1:500
Cyclin-D1	Rabbit IgG	SP4	NeoMarkers, Fremont	RM-9104-R7	Heat induced	1:100
Cyclin-E	Mouse monoclonal	13A3	NeoMarkers	MS-1060s1	Citrate buffer (pH, 6)	1:10
E-CAD	Mouse monoclonal	HECD-1	Zymed Laboratories	08-1222	Citrate buffer (pH, 6)	1:4
EGFR	Mouse monoclonal	2-18C9	Dako	K1492	Proteinase K	Ready
ER	Rabbit monoclonal	SP1	Lab Vision Corp	RM9101	Citrate buffer (pH, 6)	1:200
Galectin-3	IgG1	9C4	Vector Laboratories, Burlingame, California	VP-6802	S20 EDTA	1:250
HBME-1	IgMκ	HBME-1	Dako	M3505	PC8 citrate	1:50
HER2	Rabbit polyclonal	NA	Dako	A485	Steam, 20 min, TRS	1:500
HER3	Rabbit polyclonal	NA	NeoMarkers	RB-066-PO	None	1:200
HER4	Mouse monoclonal	HFR1	NeoMarkers	MS-637-PO	None	1:160
HSP-27	Mouse monoclonal IgG	G31	Stressgen, San Diego, California	SPA 800	Heat induced	1:100
IGF1-R	Rabbit polyclonal	NA	Santa Cruz Biotechnology	SC-713	Citrate buffer (pH, 6)	1:100
IGFBP2	Goat polyclonal	NA	Santa Cruz Biotechnology	NA	CC1(EDTA buffer) VENTANA	1:50
IGFBP5	Goat polyclonal	NA	Santa Cruz Biotechnology	NA	CC1(EDTA buffer) VENTANA	1:25
ILK	Mouse monoclonal IgG	65.1	Santa Cruz Biotechnology	SC-20019	Heat induced	1:20
INH	Mouse IgG2a	R1	Oxford Bio Innovation, Oxfordshire, England	MCA951S	Heat induced	1:50
KI67	Rabbit polyclonal	NA	Lab Vision Corp	NA	Heat induced	1:200
MDM2	Mouse monoclonal IgG	SMP14	NeoMarkers	MS-291-PO	Heat induced	1:400
MIB-1	Rabbit monoclonal IgG	SMP14	NeoMarkers	MS-291-PO	Heat induced	1:200
MLH1	IgGκ	G168-15	Biocare, Concord, California	CM220C	S30 EDTA	1:10
MSH2	IgGκ	FE11	Zymed Laboratories	33-7900	PC8 citrate	1:100
MSH6	IgG1	44	BD Biosciences, San Jose, California	610919	PC8 citrate	1:1000
O13	Mouse monoclonal IgG1	O13	ID Laboratories, Cambridge, Massachusetts	BP703	Heat induced	1:20
P16	Mouse monoclonal IgG1κ	16PO4	Cell Marque, Rocklin, California	CMA 800	Heat induced	1:20
P21	Mouse monoclonal IgG	Dcs-60.2	NeoMarkers	MS-230-PO	Heat induced	1:20

Antibody name	Isotype	Clone	Company	Catalog No.	Antigen Retrieval	Conc.
P27	Mouse monoclonal IgG	Dcs-72.f6	NeoMarkers	MS-256-PO	Citrate buffer (pH, 6)	1:100
P504S	Rabbit IgG	13H4	Zeta, Kesselsdorf, Germany	Z2001	PC8 citrate	1:50
P53	Mouse IgG2bκ	DO-7	Dako	M7001	Heat induced	1:400
P57	Mouse monoclonal	57PO6	Lab Vision Corp	MS-1062-PO	Heat induced	1:50
P63	Mouse IgG2aκ	4A4	Cell Marque	CMC441R	Heat induced	1:200
P75-NTR	IgGκ	NGFR5	NeoMarkers	MS-394-P	S20Cit+10' pronase	1:1000
P-AKT	Rabbit monoclonal	Ser 473	Cell Signalling Technology, Danvers, Massachusetts	3787	Heat induced	1:2
PGI	Rabbit polyclonal	NA	R. Nabi, University of British Columbia	NA	Heat induced	1:25
PMS2	Mouse monoclonal IgG	A16-4	BD Pharmingen, Franklin Lakes, New Jersey	556415	S30 EDTA	1:50
PR	Rabbit monoclonal	SP2	Lab Vision Corp	RM9102	Citrate buffer (pH, 6)	1:400
PSA	Mouse monoclonal	ER-PR8	Dako	M0750	Heat induced	1:100
RET	Mouse monoclonal	3F8	NeoMarkers	MS-1120-S1	Heat induced	1:2
S100	Rabbit	NA	University of Toronto, Toronto, Ontario, Canada	NA	None	1:2000
Syntrophin	Mouse monoclonal	1351	American BioReagents, Golden, Colorado	MAI-745	Citrate buffer (pH, 6)	1:100
TDT	Rabbit polyclonal	NA	Dako	A3524	Heat induced	1:20
TG	Rabbit polyclonal	NA	Dako	A0251	Heat induced	1:10 000
TOPO-II	IgG1κ	SWT3D1	Oncogene, Cambridge, Massachusetts	NA-14-100UG	PC8 citrate	1:800
TS106	IgG1	Ts106	Chemicon International, Temecula, California	MAB4130	PC15 citrate	1:100
TSH	Rabbit anti-human	NA	Dako	A0574	Heat induced	1:500
TTF-1	Mouse monoclonal	8G7G3/1	Dako	M3575	Heat induced	1:100
UPA-R	Mouse monoclonal IgG	NA	American Diagnostica, Inc, Montreal, Quebec	3689	Protease	1:200
VEGF	Mouse monoclonal	26503.11	R&D Systems, Minneapolis, Minnesota	MAB 293	Citrate buffer (pH, 6)	1:500
WT1	Mouse monoclonal	6F-H2	Dako	M3561	Heat induced	1:100

Table 4.5. Scoring System Types for Markers Evaluated

Scoring Type	Scoring System
A	1 = Herceptest positive (Dako, Glostrup, Denmark)
	0 = Herceptest negative (Dako)
B	1 = Positive (>=5% of cells)
	0 = Negative (<5% of cells)
C	2 = Strong
	1 = Weak
	0 = Negative
D	2 = Diffuse strong
	1 = Focal strong or diffuse weak
	0 = Focal weak or negative
E	2 = >50% Cells positive
	1 = 5%-50% Cells positive
	0 = <5% Cells positive
F	3 = >50% Cells positive
	2 = 26%-50% Cells positive
	1 = 5%-25% Cells positive
	0 = <5% Cells positive
G	3 = >75% Of cells positive
	2 = 26%-75% Of cells positive
	1 = 5%-25% Of cells positive
	0 = <5% Of cells positive
H	3 = >50% Of cells positive
	2 = 11%-50% Of cells positive
	1 = 5%-10% Of cells positive
	0 = <5% Of cells positive
I	3 = Strong
	2 = Moderate
	1 = Weak
	0 = Negative
J	1 = Expression retained
	0 = Loss of expression
K	3 = Moderate/strong signal in >75% of cells
	2 = Moderate/strong signal in 10%-75% of cells
	1 = Weak signal in >50% of cells
	0 = No signal or weak signal in <50% of cells

Table 4.6. Comparison groups analyzed for overlap

Each overlap analysis group defines an artificial group of comparisons for which gene overlap was assessed. To illustrate, the 'cancer vs. non-cancer' group includes all comparisons between what we would consider 'cancer' (as defined in condition set 1) and 'non-cancer' (as defined in condition set 2). In this case, 21 comparisons were identified from the literature that met the criteria and produced a list of 755 potential cancer markers. A total of 107 genes were identified in multiple studies with a consistent direction of change (18 additional genes were identified in multiple studies but with inconsistent direction of change). This result was found to be significant by Monte Carlo simulation ($p<0.0001$). The phrase 'Any other' refers to the fact that some comparison groups are comprised of all comparisons that contrasted any sample in condition set 1 to any sample not in condition set 1. For example, 'papillary cancer vs. other' included all comparisons of papillary cancers (PTC, FVPTC, PCL) to any other kind of sample (normal, cancer or benign) as long as it was not papillary. In this table, numbers are for mapped genes. All genes in the cancer/normal and cancer/benign subgroups also appear in the cancer/non-cancer group. However, genes from one comparison appear in cancer/non-cancer but neither of the subgroups because the study compared FTC to both FA and normal samples grouped together (i.e. neither normal nor benign alone).

Overlap analysis group	**Condition set 1**	**Condition set 2**	**No. comparisons**	**No. genes (with multi-study confirmation)**	**p-value**
Cancer vs. non-cancer	ACL, ATC, FCL, FTC, FVPTC, HCC, M, MACL, PCL, PTC, TCVPTC, UCL	AFTN, CTN, FA, GT, HN, Norm	21	755 (107)	<0.0001
Cancer vs. normal	ACL, ATC, FCL, FTC, FVPTC, HCC, M, MACL, PCL, PTC, TCVPTC, UCL	Norm	12	478 (53)	<0.0001
Cancer vs. benign	ACL, ATC, FCL, FTC, FVPTC, HCC, M, MACL, PCL, PTC, TCVPTC, UCL	AFTN, CTN, FA, GT, HN	8	332 (38)	<0.0001
Normal vs. benign	Norm	AFTN, CTN, FA, GT, HN	3	19 (1)	0.0113
Papillary cancer vs. non-cancer	FVPTC, PCL, PTC, TCVPTC	AFTN, CTN, FA, GT, HN, Norm	12	503 (82)	<0.0001
Papillary cancer vs. normal	FVPTC, PCL, PTC, TCVPTC	Norm	8	369 (49)	<0.0001
Papillary cancer vs. benign	FVPTC, PCL, PTC, TCVPTC	AFTN, CTN, FA, GT, HN	4	183 (13)	<0.0001
Papillary cancer vs. other	FVPTC, PCL, PTC, TCVPTC	Any other	15	528 (107)	<0.0001
FVPTC vs. other	FVPTC	Any other	2	157 (0)	N/A
FTC vs. FA	FTC	FA	6	222 (3)	0.0455
Follicular cancer vs. other	FTC, FCL	Any other	10	403 (15)	0.0003
Aggressive cancer vs. other	ACL, ATC, M, MACL	Any other	4	145 (4)	0.0402
Anaplastic cancer vs. other	ACL, ATC, MACL	Any other	3	91 (6)	<0.0001
Cancer vs. non-cancer (reanalyzed Affymetrix subset)	PTC, FTC	Norm, FA	5	1317 (179)	<0.0001

Table 4.7. Cancer versus non-cancer multi-study genes

The table presents a partial list (genes identified in 3 or more comparisons) from the cancer vs. non-cancer analysis. Complete lists for this group and all others are available as supplementary data (www.bcgsc.ca/bioinfo/ge/thyroid/). * indicates a gene found in the 'cancer vs. normal' comparison group, ** 'cancer vs. benign', and *** both groups. 'Comps' refers to the number of comparisons identifying the same gene as differentially expressed (with consistent direction of change) and 'N' refers to the total number of samples used. For both 'Comps' and 'N' the number for which fold change (FC) values were available is indicated in brackets. 'Mean FC' represents the average of all reported fold changes for that gene. To illustrate, MET was identified as consistently 'up-regulated' in six independent comparisons totalling 202 patient samples. Of these, four comparisons with a total of 162 samples reported fold change values resulting in an average fold-change of 4.55. The column labelled 'Refs' lists references for the actual studies which originally reported the gene as differentially expressed.

Gene	Description	Comps Up/Down (with FC)	N (with FC)	Mean FC (Range)	Refs
MET***	met proto-oncogene (hepatocyte growth factor receptor)	6/0 (4)	202 (162)	4.54 (2.60 to 6.60)	[20, 26, 27, 32, 64, 73]
TFF3***	trefoil factor 3 (intestinal)	0/6 (4)	196 (146)	-22.04 (-63.55 to -3.80)	[18, 20, 26, 32, 60, 73]
SERPINA1***	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	6/0 (6)	192 (192)	15.84 (5.00 to 27.64)	[20, 26, 27, 32, 60]
EPS8***	epidermal growth factor receptor pathway substrate 8	5/0 (5)	186 (186)	3.14 (2.10 to 3.80)	[20, 26, 27, 32, 34]
TIMP1***	tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibitor)	5/0 (5)	142 (142)	5.37 (3.20 to 10.31)	[26, 27, 60, 64]
TGFA***	transforming growth factor, alpha	4/0 (3)	165 (146)	6.18 (3.20 to 7.91)	[20, 26, 27, 73]
QPCT***	glutaminyl-peptide cyclotransferase (glutaminyl cyclase)	4/0 (4)	153 (153)	7.31 (3.40 to 11.67)	[20, 26, 27, 29]
PROS1***	protein S (alpha)	4/0 (3)	149 (130)	5.76 (3.40 to 7.39)	[20, 26, 32, 73]
CRABP1***	cellular retinoic acid binding protein 1	0/4 (4)	146 (146)	-11.54 (-24.45 to -2.20)	[20, 26, 32, 60]
FN1***	fibronectin 1	4/0 (4)	128 (128)	7.67 (5.20 to 10.30)	[26, 27, 32, 64]
FCGBP***	Fc fragment of IgG binding protein	0/4 (3)	108 (89)	-3.20 (-3.30 to -3.10)	[26, 60, 73]
TPO***	thyroid peroxidase	0/4 (3)	91 (89)	-6.25 (-8.60 to -2.70)	[26, 32, 60, 136]
LRP4***	low density lipoprotein receptor-related protein 4	3/0 (3)	146 (146)	14.47 (6.40 to 19.43)	[20, 26, 27]
PSD3***	pleckstrin and Sec7 domain containing 3	3/0 (3)	146 (146)	3.99 (2.70 to 5.50)	[20, 26, 27]
C11orf8***	chromosome 11 open reading frame 8	0/3 (3)	134 (134)	-7.04 (-12.49 to -2.25)	[20, 32, 134]
FABP4***	fatty acid binding protein 4, adipocyte	0/3 (3)	130 (130)	-8.55 (-15.36 to -4.90)	[20, 26, 32]
RGS16***	regulator of G-protein signalling 16	0/3 (3)	130 (130)	-4.01 (-6.75 to -2.00)	[20, 26, 32]
SDC4***	syndecan 4 (amphiglycan, ryudocan)	3/0 (3)	130 (130)	3.32 (2.30 to 4.17)	[20, 26, 32]
COL9A3***	collagen, type IX, alpha 3	0/3 (3)	128 (128)	-13.97 (-27.39 to -4.50)	[20, 26, 60]
HBB*	hemoglobin, beta	0/3 (2)	118 (87)	-7.58 (-11.39 to -3.77)	[18, 20, 27]
ETV5***	ets variant gene 5 (ets-related molecule)	3/0 (3)	111 (111)	3.60 (2.98 to 4.38)	[20, 27, 34]
CD44***	CD44 antigen (homing function and Indian blood group system)	3/0 (3)	111 (111)	3.12 (2.24 to 4.51)	[20, 27, 34]
FCGRT***	Fc fragment of IgG, receptor, transporter, alpha	0/3 (1)	109 (59)	-2.8 (-2.80 to -2.80)	[18, 26, 73]
CITED1***	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 1	3/0 (3)	107 (107)	18.73 (7.90 to 26.90)	[26, 27, 32]
KRT19***	keratin 19	3/0 (3)	107 (107)	6.55 (4.00 to 9.35)	[26, 27, 32]
GPR51***	G protein-coupled receptor 51	3/0 (3)	107 (107)	5.67 (3.30 to 8.26)	[26, 27, 60]
LGALS3***	lectin, galactoside-binding, soluble, 3 (galectin 3)	3/0 (3)	107 (107)	3.7 (3.50 to 3.80)	[26, 27, 32]
DPP4*	dipeptidylpeptidase 4 (CD26, adenosine deaminase complexing protein 2)	3/0 (3)	103 (103)	46.19 (8.20 to 115.76)	[20, 27, 32]
TUSC3*	tumour suppressor candidate 3	3/0 (3)	103 (103)	5.84 (2.43 to 7.70)	[20, 27, 32]
P4HA2*	procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), alpha polypeptide II	3/0 (3)	103 (103)	3.75 (2.93 to 4.50)	[20, 27, 32]
CCND1*	cyclin D1 (PRAD1: parathyroid adenomatosis 1)	3/0 (2)	101 (87)	2.93 (2.49 to 3.37)	[20, 27, 30]
DIO1***	deiodinase, iodothyronine, type I	0/3 (2)	94 (75)	-3.75 (-5.20 to -2.30)	[26, 32, 73]
ITPR1***	inositol 1,4,5-triphosphate receptor, type 1	0/3 (2)	94 (75)	-2.6 (-2.70 to -2.50)	[26, 32, 73]
MT1F**	metallothionein 1F (functional)	0/3 (2)	92 (73)	-2.85 (-2.91 to -2.80)	[26, 60, 73]

Gene	Description	Comps Up/Down (with FC)	N (with FC)	Mean FC (Range)	Refs
PHLDA2***	pleckstrin homology-like domain, family A, member 2	3/0 (3)	89 (89)	8.02 (2.50 to 15.96)	[26, 32, 60]
MT1G***	metallothionein 1G	0/3 (3)	89 (89)	-5.55 (-8.60 to -2.30)	[26, 32, 60]
ID4**	inhibitor of DNA binding 4, dominant negative helix-loop-helix protein	0/3 (2)	89 (70)	-3.64 (-5.00 to -2.29)	[29, 73, 134]
DUSP6***	dual specificity phosphatase 6	3/0 (3)	72 (72)	4.5 (3.68 to 5.22)	[27, 32, 34]
HBA2*	hemoglobin, alpha 2	0/3 (3)	69 (69)	-3.5 (-4.72 to -2.38)	[27, 32, 64]

Table 4.8. Gene Ontology analysis of multi-study genes from the cancer versus non-cancer overlap analysis group

Of the 107 genes with multi-study confirmation from the cancer versus non-cancer overlap analysis group, 102 were present in the Gene Ontology (GO) set of 15,240 human genes. From this list, a total of 12 GO terms were found to be statistically over-represented: 3 biological process (P), 3 cellular component (C), and 6 molecular function (F) terms. The 'obs/total' column shows the number of genes from the list found associated with each GO term over the total number of genes annotated to that term in GO. The p-value was calculated using the hypergeometric test, corrected with a Benjamini & Hochberg False Discovery Rate (FDR) correction and a cut off of 0.05 applied to the result.

GO term (GO ID)	Ontology	obs/total	p-value	Genes in test set
extracellular region (5576)	C	23/1,093	4.09E-04	CYR61, CHI3L1, TIMP1, TNFRSF11B, ADM, GPC3, LOX, PLAU, TFF3, LGALS3, TGFA, CCL21, SPINT1, SERPINA1, RNASE1, PROS1, TNC, DPT, MATN2, IGFBP3, COL9A3, BMP1, FN1
cadmium ion binding (46870)	F	3/4	4.09E-04	MT1F, MT1E, MT1A
thyroid hormone generation (6590)	P	3/5	6.78E-04	DIO2, DIO1, TPO
thyroid hormone metabolism (42403)	P	3/6	1.01E-03	DIO2, DIO1, TPO
selenium binding (8430)	F	3/7	1.41E-03	DIO2, SELENBP1, DIO1
hormone metabolism (42445)	P	5/48	1.97E-03	HSD17B6, DIO2, DIO1, TPO, ADM
MAP kinase phosphatase activity (17017)	F	3/10	3.40E-03	DUSP6, DUSP4, DUSP1
thyroxine 5'-deiodinase activity (4800)	F	2/3	1.16E-02	DIO2, DIO1
copper ion binding (5507)	F	4/41	1.25E-02	MT1F, MT1E, MT1A, LOX
extracellular matrix (sensu Metazoa) (5578)	C	9/342	3.45E-02	CHI3L1, TIMP1, TNC, DPT, MATN2, COL9A3, GPC3, LOX, FN1
extracellular matrix (31012)	C	9/347	3.49E-02	CHI3L1, TIMP1, TNC, DPT, MATN2, COL9A3, GPC3, LOX, FN1
retinoic acid receptor activity (3708)	F	2/6	3.82E-02	RXRG, RARA

Table 4.9. Summary of marker staining in malignant versus benign array for first score grouping and ungrouped data

In the first score grouping, marker scores were grouped as either negative (score=0) or positive (score≥1). The table shows the number of patient samples staining negative or positive and the percent positive for benign versus malignant tumours. 'Cont. table' refers to contingency table statistics (Pearson χ^2 or Fisher exact test where appropriate). BH refers to the Benjamini & Hochberg multiple testing correction. MU refers to the Mann-Whitney U test. "Var. Imp." refers to the standard variable importance assigned by the Random Forests (RF) classifier. For contingency table statistics, marker scores were grouped. For MU test and RF classification, ungrouped data were used. 'Direction' refers to the direction of change, "up-regulated" or "down-regulated", for the marker expression in malignant samples relative to benign samples (determined by difference in mean rank in the MU test). For example, VEGF appears to be down-regulated with decreased expression in malignant samples compared to malignant. Only significant results (by MU test) are summarized in the table. *Indicates markers which were also identified in the meta-analysis (with an overlap of at least 3).

	Benign			Malignant			Statistical Tests			
Marker	**Positive no.**	**Negative no.**	**Positive %**	**Positive no.**	**Negative no.**	**Positive %**	**Cont. table P-value (BH)**	**MU P-value (BH)**	**Var. Imp.**	**Direction**
VEGF	90	10	90	36	60	37.5	0	0.0000	6.2	Down
Galectin-3*	29	71	29	87	9	90.6	0	0.0000	14.2	Up
CK19*	67	33	67	95	1	99	0	0.0000	12.9	Up
AR	2	98	2	50	46	52.1	0	0.0000	5.3	Up
Aurora-A	20	78	20.4	67	23	74.4	0	0.0000	5.6	Up
HBME-1	5	95	5	53	43	55.2	0	0.0000	4.4	Up
P16	5	95	5	52	43	54.7	0	0.0000	4.3	Up
Bcl-2	78	22	78	32	61	34.4	0	0.0000	2.3	Down
Cyclin-D1*	43	51	45.7	74	11	87.1	0	0.0000	3.2	Up
CAV-1	10	89	10.1	50	46	52.1	0	0.0000	2.2	Up
E-CAD	100	0	100	94	2	97.9	0.3612	0.0000	3.0	Down
Cyclin-E	37	62	37.4	68	26	72.3	0	0.0000	2.1	Up
Clusterin	25	74	25.3	61	35	63.5	0	0.0000	1.3	Up
CR3	17	81	17.3	50	41	54.9	0	0.0000	2.3	Up
IGFBP5	17	81	17.3	48	43	52.7	0	0.0000	1.7	Up
P21	38	62	38	66	26	71.7	0	0.0000	0.6	Up
CTNNB1	1	99	1	19	77	19.8	0.0001	0.0000	0.2	Up
IGFBP2	16	82	16.3	43	49	46.7	0	0.0001	1.8	Up
Caveolin	60	38	61.2	69	18	79.3	0.0249	0.0002	2.2	Up
HER4	65	33	66.3	84	12	87.5	0.0024	0.0003	1.5	Up
TG	97	0	100	94	0	100	NA	0.0003	2.5	Down
c-kit	19	81	19	2	91	2.2	0.0013	0.0004	0.8	Down
S100	0	99	0	12	78	13.3	0.0016	0.0004	0.2	Up
KI67	2	95	2.1	16	74	17.8	0.0019	0.0007	0.8	Up
Aurora-C	52	44	54.2	66	20	76.7	0.006	0.0015	1.7	Up
RET	13	72	15.3	35	57	38	0.0031	0.0017	0.7	Up
HER3	34	65	34.3	48	37	56.5	0.0098	0.0056	0.7	Up
AMF-R	20	78	20.4	36	57	38.7	0.019	0.0113	0.8	Up
MLH1	99	0	100	89	7	92.7	0.014	0.0124	1.3	Down
TTF-1	83	15	84.7	84	7	92.3	0.2721	0.0149	1.1	Up
AAT*	7	92	7.1	17	71	19.3	0.0434	0.0194	1.0	Up
Syntrophin	18	79	18.6	32	62	34	0.0434	0.0267	0.7	Up
HSP-27	88	11	88.9	75	9	89.3	1	0.0351	1.7	Down

Table 4.10. Summary of marker staining in malignant versus benign array for second score grouping

In the second score grouping, marker scores were grouped as either negative/low (score≤1) or positive/high (score≥2). The table shows the number of patient samples staining negative or positive and the percent positive for benign versus malignant tumours. 'Cont. table' refers to contingency table statistics (Pearson χ^2 or Fisher exact test where appropriate). BH refers to the Benjamini & Hochberg multiple testing correction. 'Direction' refers to the direction of change, "up-regulated" or "down-regulated", for the marker expression in malignant samples relative to benign samples (determined by change in Positive % for contingency table statistic). Only significant results are summarized in the table. * Indicates markers which were also identified in the meta-analysis (with an overlap of at least 3).

	Benign			Malignant			Statistical Tests	
Marker	**Positive no.**	**Negative no.**	**Positive %**	**Positive no.**	**Negative no.**	**Positive %**	**Cont. table P-value (BH)**	**Direction**
Galectin-3*	87	13	13.0	12	84	87.5	0.0000	Up
CK19*	54	46	46.0	3	93	96.9	0.0000	Up
VEGF	30	70	70.0	81	15	15.6	0.0000	Down
E-CAD	29	71	71.0	71	25	26.0	0.0000	Down
Cyclin-D1*	67	27	28.7	22	63	74.1	0.0000	Up
Aurora-A	95	3	3.1	60	30	33.3	0.0000	Up
P16	100	0	0.0	71	24	25.3	0.0000	Up
Cyclin-E	83	16	16.2	46	48	51.1	0.0000	Up
Bcl-2	71	29	29.0	91	2	2.2	0.0000	Down
CR3	86	12	12.2	54	37	40.7	0.0001	Up
CAV-1	95	4	4.0	71	25	26.0	0.0002	Up
TG	2	95	97.9	18	76	80.9	0.0014	Down
IGFBP5	89	9	9.2	64	27	29.7	0.0029	Up
HER4	69	29	29.6	44	52	54.2	0.0035	Up
Caveolin	59	39	39.8	34	53	60.9	0.0243	Up

Table 4.11. Summary of marker staining in matched ATC versus DTC for first score grouping and ungrouped statistics

In the first score grouping, marker scores were grouped as either negative (score=0) or positive (score≥1). The table shows the number of patient samples staining negative or positive and the percent positive for patient-matched ATC versus DTC. 'Cont. table' refers to contingency table statistics. BH refers to the Benjamini & Hochberg multiple testing correction. MH refers to the marginal homogeneity test. "Var. Imp." refers to the variable importance assigned by the RF classifier. For contingency table statistics, marker scores were grouped. For MH test and RF classification, ungrouped data were used. 'Direction' refers to the direction of change, "up-regulated" or "down-regulated", for the marker expression in ATC samples relative to DTC samples (determined by change in Positive % for contingency table statistic). For example, MIB-1 appears to be up-regulated with increased expression in ATC samples compared to DTC. Only results for significant markers are summarized in the table.

	DTC			ATC			Statistical Tests			
Marker	**Positive no.**	**Negative no.**	**Positive %**	**Positive no.**	**Negative no.**	**Positive %**	**Cont. table P-value (BH)**	**MH P-value (BH)**	**Var. Imp.**	**Direction**
MIB-1	3	9	25.0	12	0	100.0	0.005	0.032	15.6	Up
VEGF	11	1	91.7	6	6	50.0	0.300	0.032	5.3	Down
E-CAD	11	1	91.7	2	10	16.7	0.008	0.032	9.6	Down
TG	11	1	91.7	1	10	9.1	0.004	0.032	17.3	Down
CTNNB1	11	1	91.7	5	7	41.7	0.186	0.032	13.0	Down
Bcl-2	10	2	83.3	1	11	8.3	0.005	0.035	12.4	Down
P53	2	10	16.7	10	2	83.3	0.010	0.040	5.8	Up
TOPO-II	5	7	41.7	11	1	91.7	0.186	0.040	5.0	Up

Table 4.12. Summary of marker staining in matched ATC versus DTC for second score grouping

In the second score grouping, marker scores were grouped as either negative/low (score≤1) or positive/high (score≥2). The table shows the number of patient samples staining negative or positive and the percent positive for patient-matched ATC versus DTC. 'Cont. table' refers to contingency table statistics. BH refers to the Benjamini & Hochberg multiple testing correction. 'Direction' refers to the direction of change, "up-regulated" or "down-regulated", for the marker expression in ATC samples relative to DTC samples (determined by change in Positive % for contingency table statistic). Only results for significant markers are summarized in the table.

	DTC			ATC			Statistical Tests	
Marker	**Positive no.**	**Negative no.**	**Positive %**	**Positive no.**	**Negative no.**	**Positive %**	**Cont. table P-value (BH)**	**Direction**
MIB-1	1	11	8.3	11	1	91.7	0.001	Up
CTNNB1	11	1	91.7	2	10	16.7	0.004	Down
TG	9	3	75.0	1	10	9.1	0.030	Down
TOPO-II	2	10	16.7	9	3	75.0	0.030	Up
VEGF	7	5	58.3	0	12	0.0	0.030	Down

Table 4.13. Summary of experimental validation for top 12 markers

The top 12 genes from the 'cancer vs. non-cancer' overlap analysis group were extensively reviewed. For each gene, experimental validation of differential expression and/or utility as a diagnostic marker in thyroid cancer is summarized at both the RNA and protein level. Abbreviations: RT-PCR, reverse transcriptase polymerase chain reaction; NB, northern blot; IHC, immunohistochemistry; IS, immunostaining; WB, western blot; ELISA, enzyme-linked immunosorbent assay. *A number of studies have implicated TPO as a useful biomarker for thyroid malignancy and were reviewed by Segev *et al.* (2003)[69]. For simplicity, we cite only the review and studies published since.

Gene	RNA	Protein
MET	RT-PCR[25, 64, 73, 139]	IHC[54, 56, 57, 140], IS[55], WB[139, 141]
TFF3	RT-PCR[25, 32, 59-62]	
SERPINA1	RT-PCR[60]	IHC[63], WB[63]
EPS8		
TIMP1	RT-PCR[60, 64], NB[142]	IHC[52, 64, 142]
TGFA	NB[70]	ELISA[70]
QPCT	RT-PCR[27]	
PROS1		
CRABP1	RT-PCR[60]	
FN1	RT-PCR[32, 64, 65]	IHC[64, 67]
FCGBP	RT-PCR[72]	
TPO	RT-PCR[30, 68, 136], NB[30, 143]	IHC[20, 69]*

Table 4.14. Evaluation of benign versus malignant TMA performance for markers identified in meta-analysis

All potential biomarkers that were both assayed in the TMA analysis and identified in the meta-analysis (even single study genes) are summarized in the table. MU refers to the Mann-Whitney U-test. BH indicates that the p-value was corrected for multiple testing by Benjamini & Hochberg method. "Var. Imp." refers to the standard variable importance assigned by the Random Forests (RF) classifier. A TMA rank out of 56 (for the 56 markers evaluated) was determined by the variable importance. 'Comps' refers to the number of comparisons identifying the same gene as differentially expressed (with consistent direction of change) and 'N' refers to the total number of samples used. For both 'Comps' and 'N' the number for which fold change (FC) values were available is indicated in brackets. 'Mean FC' represents the average of all reported fold changes for that gene. A Meta-analysis rank (Meta rank) out of 755 (for the 755 genes identified as differentially expressed in one or more profiling study) was determined by the ranking scheme described above (see meta-analysis methods). The column labelled 'Refs' lists references for the actual studies which originally reported the gene as differentially expressed.

Biomarker		TMA results			Meta-analysis results				
Antibody name	Target Gene	MU P-value (BH)	Var. Imp.	TMA Rank	Comps Up/Down (with FC)	N (with FC)	Mean FC (Range)	Meta rank	Refs
AAT	SERPINA1	0.0194	1.0	28	6/0 (6)	192 (192)	15.84 (5.00 to 27.64)	3	[20, 26, 27, 32, 60]
CK19	KRT19	0.0000	12.9	2	3/0 (3)	107 (107)	6.55 (4.00 to 9.35)	25	[26, 27, 32]
Galectin-3	LGALS3	0.0000	14.2	1	3/0 (3)	107 (107)	3.7 (3.50 to 3.80)	27	[26, 27, 32]
Cyclin-D1	CCND1	0.0000	3.2	8	3/0 (2)	101 (87)	2.93 (2.49 to 3.37)	31	[20, 27, 30]
c-kit	KIT	0.0004	0.8	31	0/2 (2)	118 (118)	-7.85 (-12.54 to -3.16)	42	[20, 134]
HER3	ERBB3	0.0056	0.7	34	1/0 (1)	59 (59)	2.1 (2.10 to 2.10)	162	[26]
TG	TG	0.0003	2.5	10	0/1 (0)	31 (0)	0 (0 to 0)	402	[18]
Bcl-2	BCL2	0.0000	2.3	11	0/1 (1)	16 (16)	-2.2 (-2.20 to -2.20)	582	[32]
IGFBP5	IGFBP5	0.0000	1.7	18	1/0 (1)	14 (14)	5.96 (5.96 to 5.96)	590	[60]
Clusterin	CLU	0.0000	1.4	21	0/1 (1)	4 (4)	-2.5 (-2.50 to -2.50)	669	[28]
E-CAD	CDH1	0.0000	3.0	9	1/1 (2)	77 (77)	0.53 (-1.74 to 2.80)	746	[30, 134]

Figure 4.1. Overlap analysis results for 'cancer vs. non-cancer' group compared to random simulation

Values shown for the 'Random permutations' bars are mean values for all permutations in the Monte Carlo simulation. A total of 755 genes were reported from 21 comparisons, and of these, 107 genes were reported more than once with a consistent fold-change direction. In three cases, genes were independently reported as many as 6 times. The total amount of overlap observed was assessed by Monte Carlo simulation. Real data was found to have significantly more overlap than simulated data ($p<0.0001$, 10,000 permutations). In 10,000 permutations, the simulated data never produced an overlap greater than three whereas real data identified 12 genes with overlap of four, five or six. The probability of observing one or more genes with an overlap of: two or more was $p=1.0$; three or more was $p=0.037$; and four or more was $p<0.0001$. The total number of genes with an overlap of two was still highly significant but we expect at least some false-positives to occur by chance in this category. Error bars were not included because standard error or 95% confidence intervals were too small to visualize.

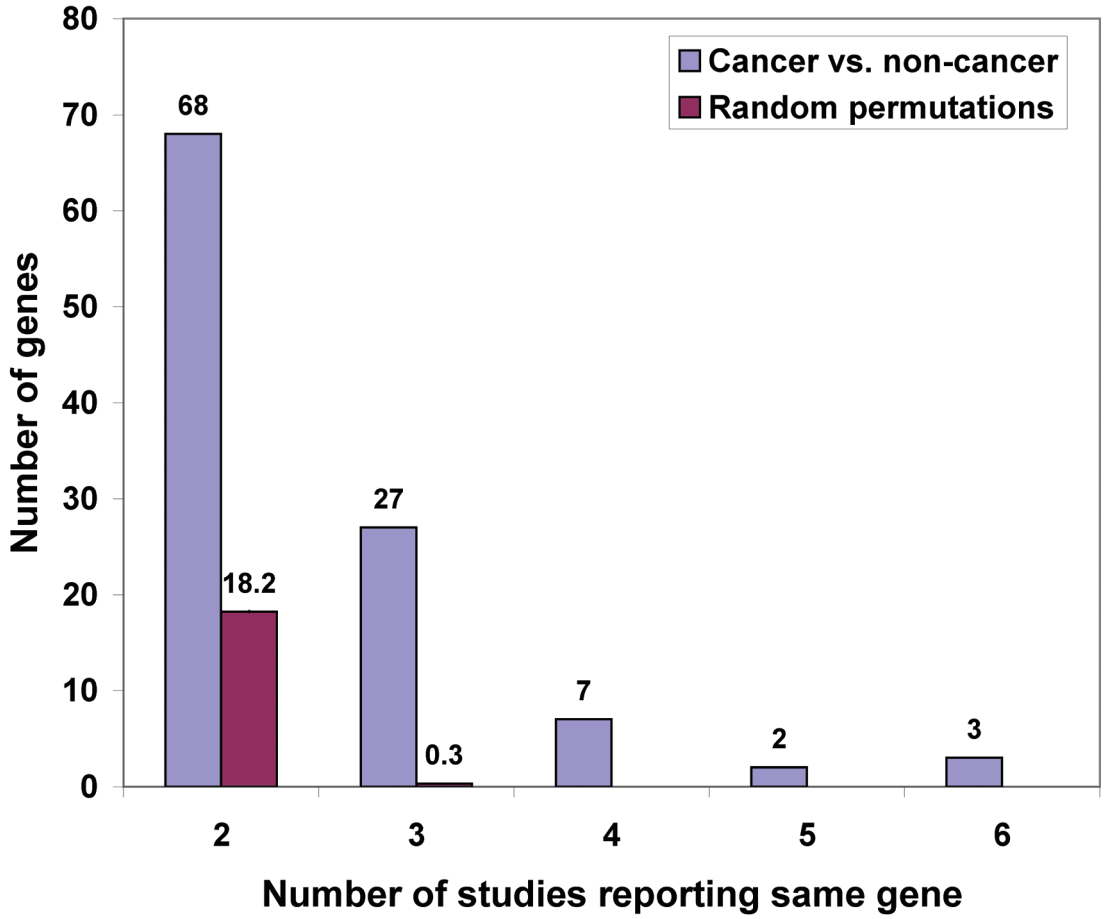


Figure 4.2. Overlap between cancer/normal and cancer/benign comparison groups

Of the 478 genes in the cancer/normal comparison group and 332 genes of the cancer/benign group, a total of 63 genes were found in both. In all, 58.9% (63/107) of the multi-study genes (two or more overlapping studies) were found in both a cancer versus normal and cancer versus benign comparison. For genes found in three or more studies, 79.5% (31/39) were reported for both types of comparisons. *, **, and *** were used to identify which part of the Venn diagram each gene in Table 4.7 corresponds to.

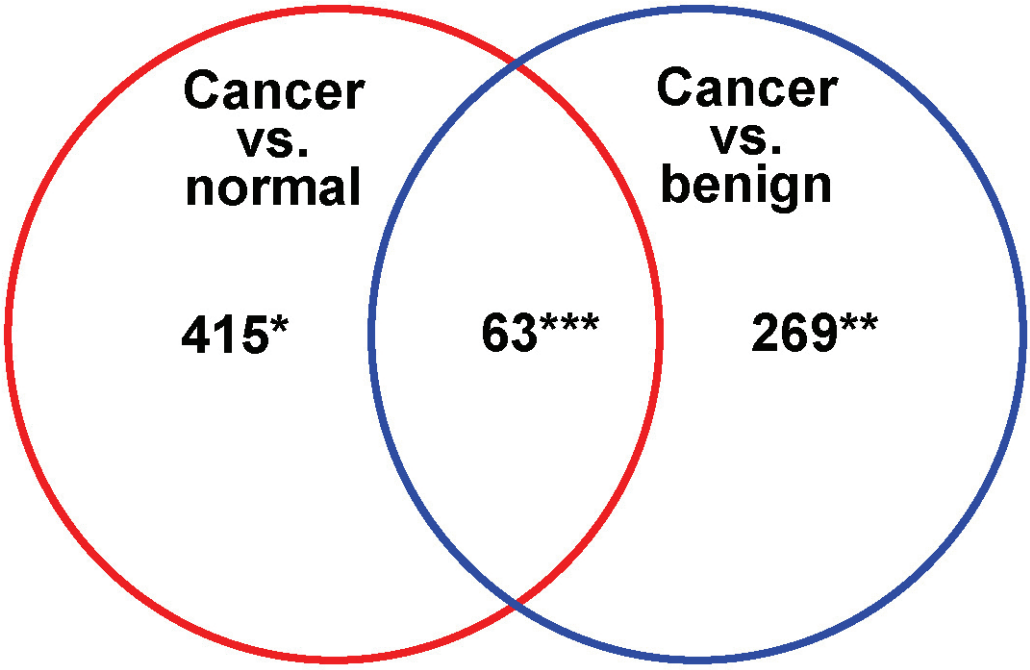


Figure 4.3. A comparison of 'cancer vs. non-cancer' genes identified with multi-study evidence based on all published lists (our meta-analysis method) versus genes identified by a smaller subset of studies re-analyzed from raw microarray data

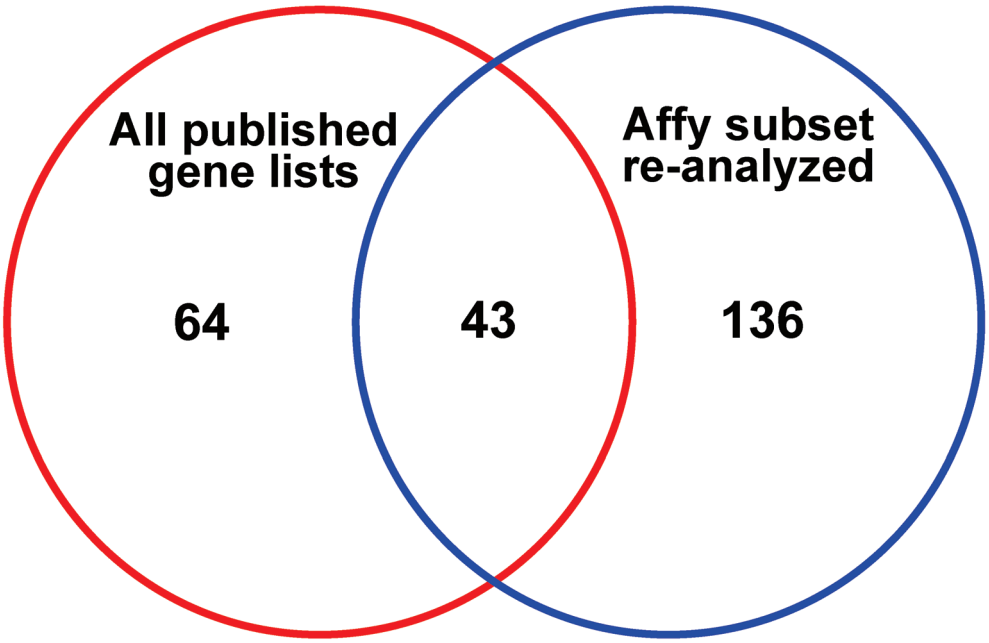


Figure 4.4. Hierarchical clustering of all 56 markers for malignant versus benign array

All 56 markers were submitted to a simple hierarchical clustering method and a heat map of marker expression was generated. Both patient samples and markers were clustered according to marker expression level. The color key indicates marker score values of 0 – 4 according to the scoring systems. Along the bottom axis, all markers are listed. Along the left axis, benign samples are indicated by a light green bar and malignant by a dark green bar. A decent separation between benign and malignant was observed but separation improved when only significant markers were included (Figure 4.5).

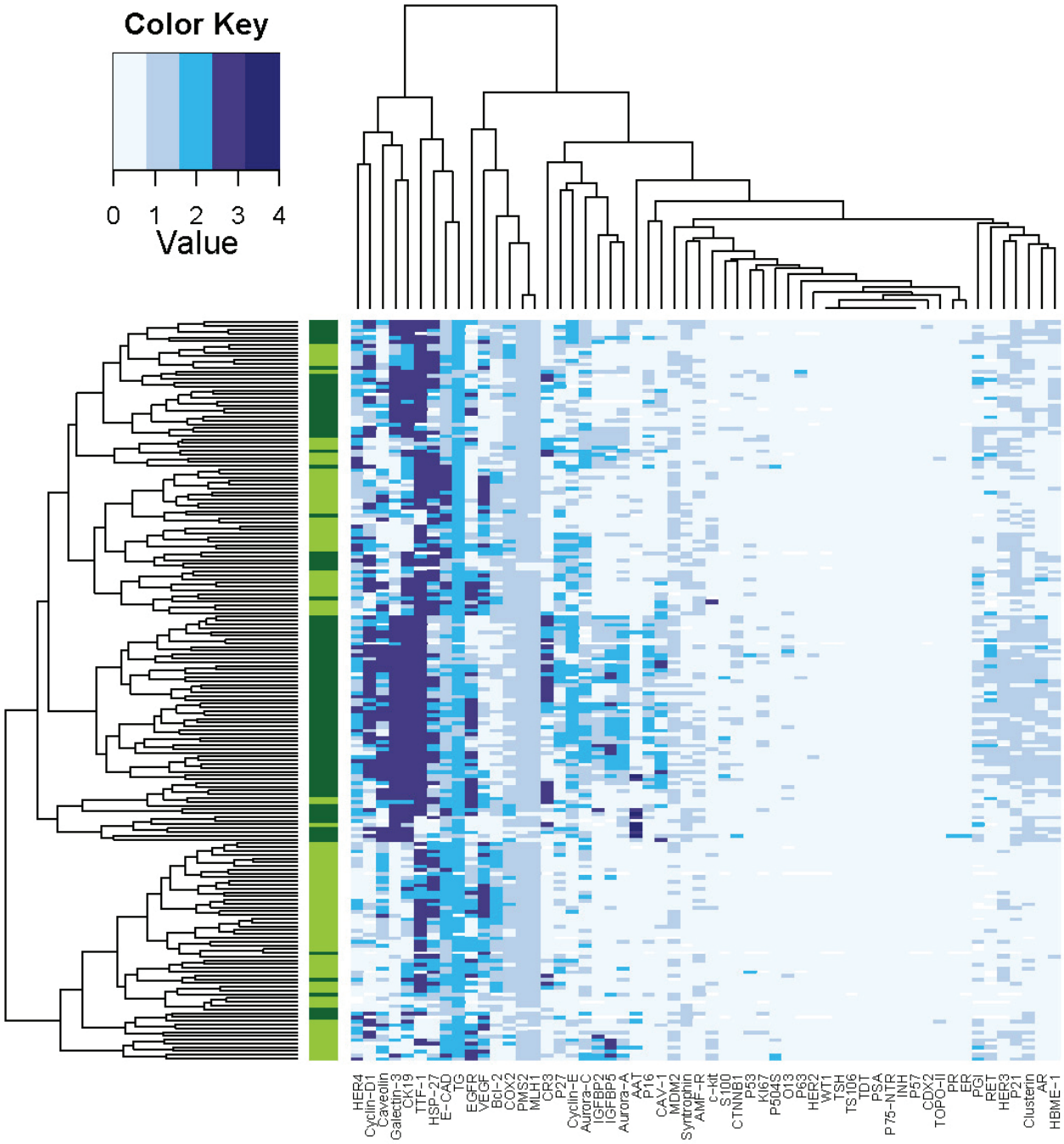


Figure 4.5. Hierarchical clustering of the ten most significant markers for malignant versus benign array

The ten most significant markers (MU test; after BH multiple testing correction) were submitted to a simple hierarchical clustering method and a heat map of marker expression was generated. Both patient samples and markers were clustered according to marker expression level. The color key indicates marker score values of 0 – 3 according to the scoring systems. Along the bottom axis, the significant markers are listed. Without the noise of the less informative markers, hierarchical clustering was better able to separate the benign samples (light green bar) from the malignant samples (dark green bar). The eight up-regulated markers (yellow bar) and two down-regulated markers (orange bar) were also clustered together.

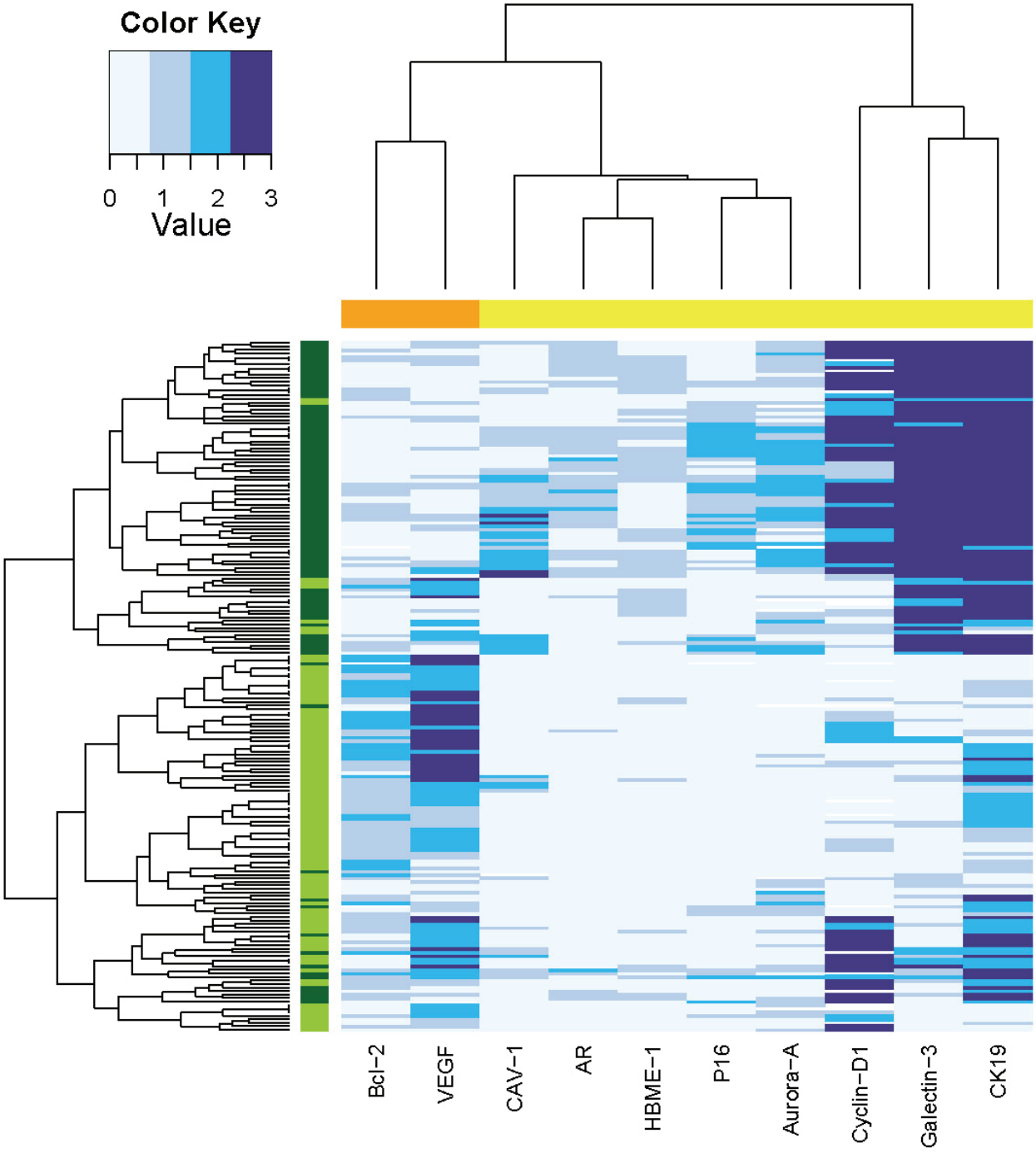


Figure 4.6. Hierarchical clustering of all 62 markers for ATC versus DTC array

All 62 markers were submitted to a simple hierarchical clustering method and a heat map of marker expression was generated. Both patient samples and markers were clustered according to marker expression level. The color key indicates marker score values of 0 – 4 according to the scoring systems. On the right axis, samples are listed with pathology (A=anaplastic, F=follicular, P=papillary) and patient number. Each of the 12 patients had two samples on the array: one ATC (light green bar: 'A' samples) and one DTC (dark green bar: 'F' or 'P' samples). Along the bottom axis, all markers are listed. A good separation between ATC and DTC was observed but separation improved when only significant markers were included (Figure 4.7).

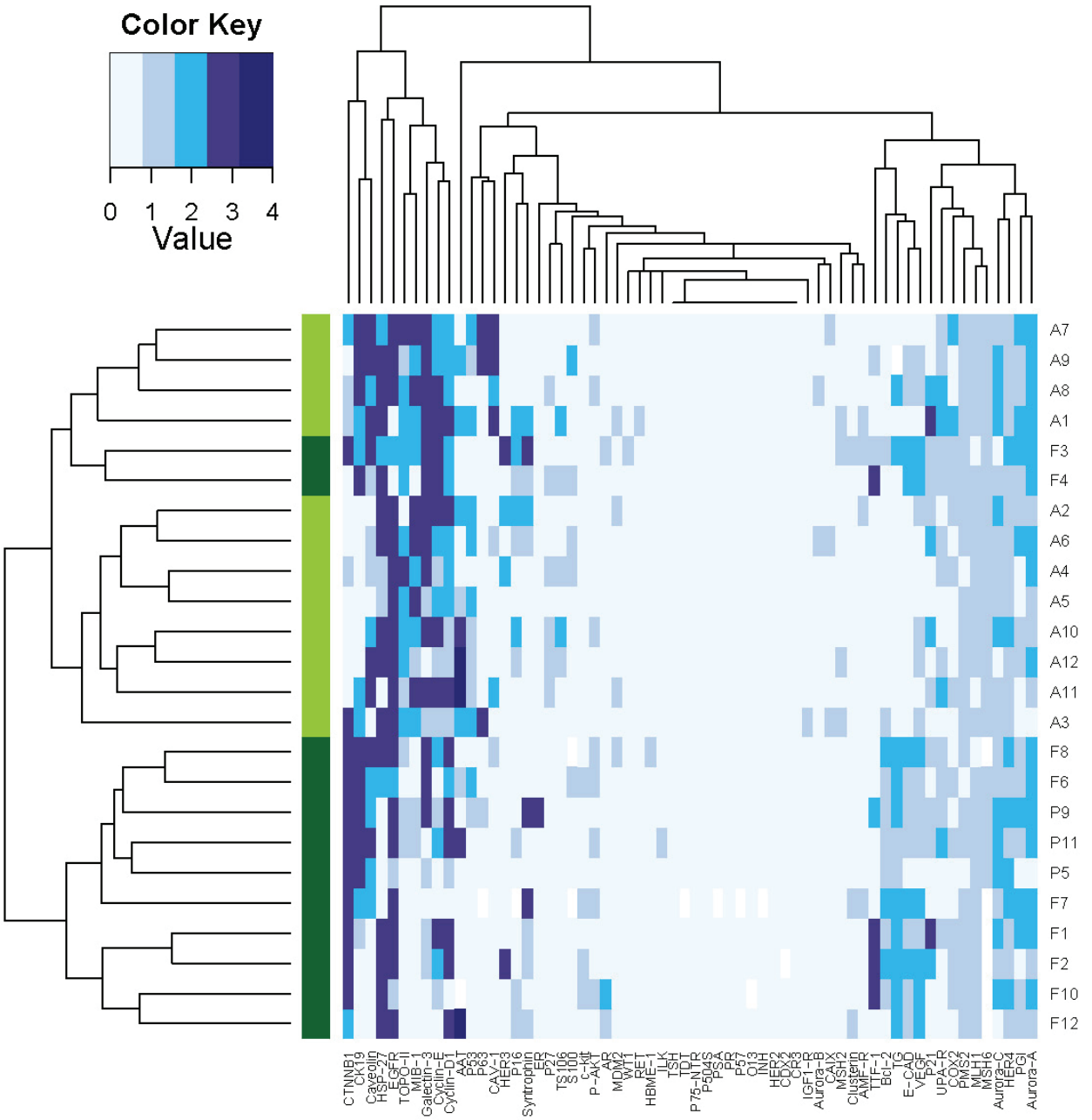
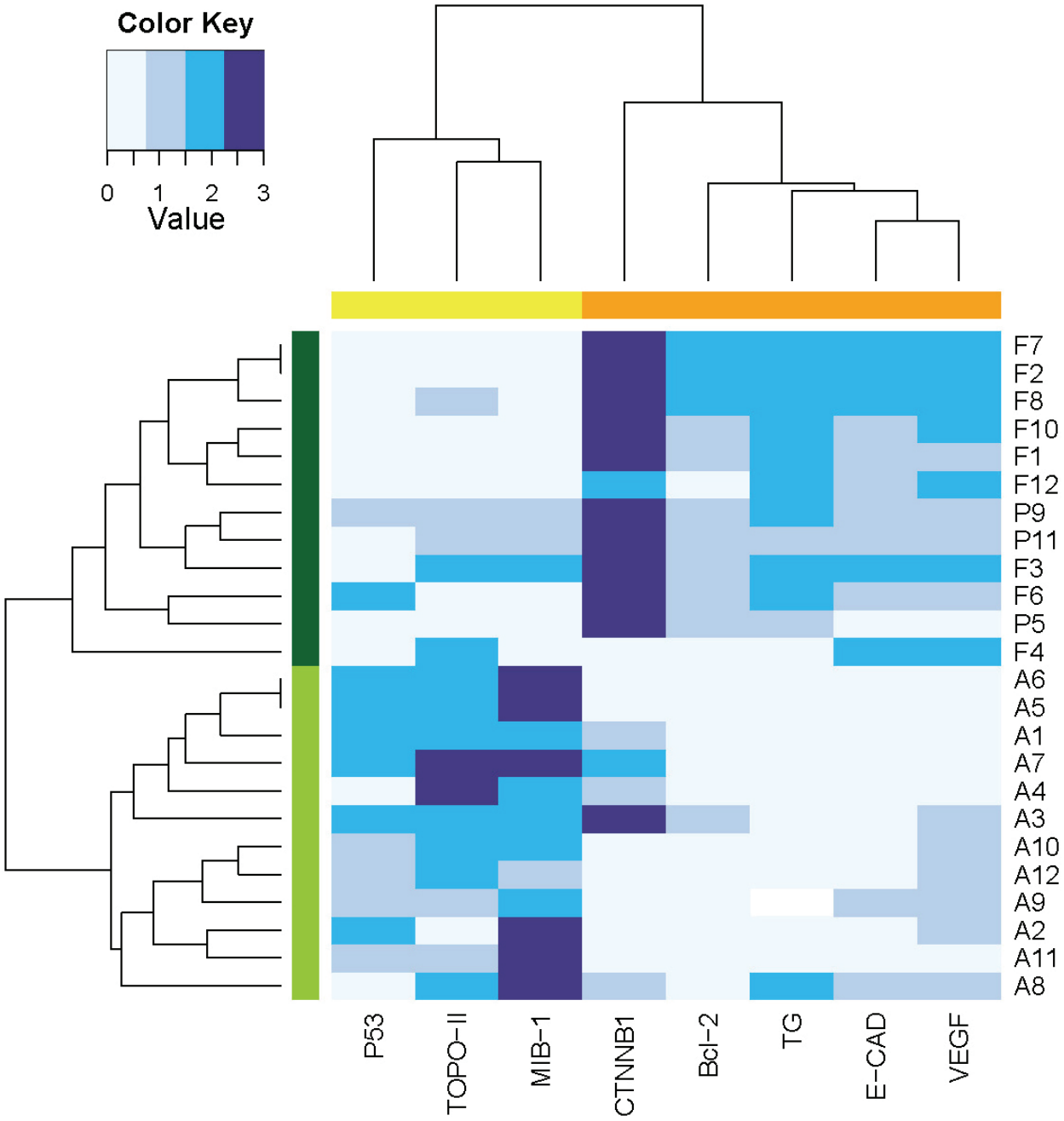


Figure 4.7. Hierarchical clustering of eight significant markers for ATC versus DTC array

The eight significant markers (MH test; after BH multiple testing correction) were submitted to a simple hierarchical clustering method and a heat map of marker expression was generated. Both patient samples and markers were clustered according to marker expression level. The color key indicates marker score values of 0 – 3 according to the scoring systems. On the right axis, samples are listed with pathology (A=anaplastic, F=follicular, P=papillary) and patient number. Each of the 12 patients had two samples on the array: one ATC (A) and one DTC (F or P). Along the bottom axis, all significant markers are listed. Without the noise of the less informative markers, hierarchical clustering was able to perfectly separate the ATC samples (light green bar) from their DTC counterparts (dark green bar). The three up-regulated markers (yellow bar) and five down-regulated markers (orange bar) were also clustered together.



References

1. Gharib, H. and J.R. Goellner, *Fine-needle aspiration biopsy of the thyroid: an appraisal.* Ann Intern Med, 1993. **118**(4): p. 282-9.
2. Greenlee, R.T., M.B. Hill-Harmon, T. Murray, and M. Thun, *Cancer statistics, 2001.* CA Cancer J Clin, 2001. **51**(1): p. 15-36.
3. Goellner, J.R., H. Gharib, C.S. Grant, and D.A. Johnson, *Fine needle aspiration cytology of the thyroid, 1980 to 1986.* Acta Cytol, 1987. **31**(5): p. 587-90.
4. Caraway, N.P., N. Sneige, and N.A. Samaan, *Diagnostic pitfalls in thyroid fine-needle aspiration: a review of 394 cases.* Diagn Cytopathol, 1993. **9**(3): p. 345-50.
5. Ravetto, C., L. Colombo, and M.E. Dottorini, *Usefulness of fine-needle aspiration in the diagnosis of thyroid carcinoma: a retrospective study in 37,895 patients.* Cancer, 2000. **90**(6): p. 357-63.
6. Wiseman, S.M., C. Baliski, R. Irvine, D. Anderson, G. Wilkins, D. Filipenko, H. Zhang, and S. Bugis, *Hemithyroidectomy: The Optimal Initial Surgical Approach for Individuals Undergoing Surgery for a Cytological Diagnosis of Follicular Neoplasm.* Ann Surg Oncol, 2006. **13**(3): p. 425-432.
7. Schena, M., D. Shalon, R.W. Davis, and P.O. Brown, *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
8. Lockhart, D.J., H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown, *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.
9. Velculescu, V.E., L. Zhang, B. Vogelstein, and K.W. Kinzler, *Serial analysis of gene expression.* Science, 1995. **270**(5235): p. 484-7.
10. Rhodes, D.R., J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey, and A.M. Chinnaiyan, *Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.* Proc Natl Acad Sci U S A, 2004. **101**(25): p. 9309-14.
11. Cahan, P., A.M. Ahmad, H. Burke, S. Fu, Y. Lai, L. Florea, N. Dharker, T. Kobrinski, P. Kale, and T.A. McCaffrey, *List of lists-annotated (LOLA): A database for annotation and comparison of published microarray gene lists.* Gene, 2005. **360**(1): p. 78-82.
12. Shih, W., R. Chetty, and M.S. Tsao, *Expression profiling by microarrays in colorectal cancer (Review).* Oncol Rep, 2005. **13**(3): p. 517-24.
13. Fox, M.S. and S. Klawansky, *Interruption of cell transformation and cancer formation.* Faseb J, 2006. **20**(13): p. 2209-13.
14. Wiseman, S.M., T.R. Loree, N.R. Rigual, W.L. Hicks, Jr., W.G. Douglas, G.R. Anderson, and D.L. Stoler, *Anaplastic transformation of thyroid cancer: review of clinical, pathologic, and molecular evidence provides new insights into disease biology and future therapy.* Head Neck, 2003. **25**(8): p. 662-70.
15. Mazzaferri, E.L. and R.T. Kloos, *Clinical review 128: Current approaches to primary therapy for papillary and follicular thyroid cancer.* J Clin Endocrinol Metab, 2001. **86**(4): p. 1447-63.
16. Shaha, A.R., J.P. Shah, and T.R. Loree, *Differentiated thyroid cancer presenting initially with distant metastasis.* Am J Surg, 1997. **174**(5): p. 474-6.
17. Gilliland, F.D., W.C. Hunt, D.M. Morris, and C.R. Key, *Prognostic factors for thyroid carcinoma. A population-based study of 15,698 cases from the Surveillance, Epidemiology and End Results (SEER) program 1973-1991.* Cancer, 1997. **79**(3): p. 564-73.

18. Onda, M., M. Emi, A. Yoshida, S. Miyamoto, J. Akaishi, S. Asaka, K. Mizutani, K. Shimizu, M. Nagahama, K. Ito, T. Tanaka, and T. Tsunoda, *Comprehensive gene expression profiling of anaplastic thyroid cancers with cDNA microarray of 25 344 genes.* Endocr Relat Cancer, 2004. **11**(4): p. 843-54.
19. Takano, T., Y. Hasegawa, F. Matsuzuka, A. Miyauchi, H. Yoshida, T. Higashiyama, K. Kuma, and N. Amino, *Gene expression profiles in thyroid carcinomas.* Br J Cancer, 2000. **83**(11): p. 1495-502.
20. Giordano, T.J., R. Kuick, D.G. Thomas, D.E. Misek, M. Vinco, D. Sanders, Z. Zhu, R. Ciampi, M. Roh, K. Shedden, P. Gauger, G. Doherty, N.W. Thompson, S. Hanash, R.J. Koenig, and Y.E. Nikiforov, *Molecular classification of papillary thyroid carcinoma: distinct BRAF, RAS, and RET/PTC mutation-specific gene expression profiles discovered by DNA microarray analysis.* Oncogene, 2005. **24**(44): p. 6646-56.
21. Pruitt, K.D., K.S. Katz, H. Sicotte, and D.R. Maglott, *Introducing RefSeq and LocusLink: curated human genome resources at the NCBI.* Trends in Genetics, 2000. **16**(1): p. 44-47.
22. Mammalian Gene Collection Program Team*, R.L. Strausberg, E.A. Feingold, L.H. Grouse, J.G. Derge, R.D. Klausner, F.S. Collins, L. Wagner, C.M. Shenmen, G.D. Schuler, S.F. Altschul, B. Zeeberg, K.H. Buetow, C.F. Schaefer, N.K. Bhat, R.F. Hopkins, H. Jordan, T. Moore, S.I. Max, J. Wang, F. Hsieh, L. Diatchenko, K. Marusina, A.A. Farmer, G.M. Rubin, L. Hong, M. Stapleton, M.B. Soares, M.F. Bonaldo, T.L. Casavant, T.E. Scheetz, M.J. Brownstein, T.B. Usdin, S. Toshiyuki, P. Carninci, C. Prange, S.S. Raha, N.A. Loquellano, G.J. Peters, R.D. Abramson, S.J. Mullahy, S.A. Bosak, P.J. McEwan, K.J. McKernan, J.A. Malek, P.H. Gunaratne, S. Richards, K.C. Worley, S. Hale, A.M. Garcia, L.J. Gay, S.W. Hulyk, D.K. Villalon, D.M. Muzny, E.J. Sodergren, X. Lu, R.A. Gibbs, J. Fahey, E. Helton, M. Ketteman, A. Madan, S. Rodrigues, A. Sanchez, M. Whiting, A. Madan, A.C. Young, Y. Shevchenko, G.G. Bouffard, R.W. Blakesley, J.W. Touchman, E.D. Green, M.C. Dickson, A.C. Rodriguez, J. Grimwood, J. Schmutz, R.M. Myers, Y.S.N. Butterfield, M.I. Krzywinski, U. Skalska, D.E. Smailus, A. Schnerch, J.E. Schein, S.J.M. Jones, and M.A. Marra, *Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences.* Proc Natl Acad Sci U S A, 2002. **99**(26): p. 16899-903.
23. Robertson, N., M. Oveisi-Fordorei, S.D. Zuyderduyn, R.J. Varhol, C. Fjell, M. Marra, S. Jones, and A. Siddiqui, *DiscoverySpace: an interactive data analysis application.* Genome Biol, 2007. **8**(1): p. R6.
24. Dennis, G., Jr., B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, and R.A. Lempicki, *DAVID: Database for Annotation, Visualization, and Integrated Discovery.* Genome Biol, 2003. **4**(5): p. P3.
25. Finley, D.J., N. Arora, B. Zhu, L. Gallagher, and T.J. Fahey, 3rd, *Molecular profiling distinguishes papillary carcinoma from benign thyroid nodules.* J Clin Endocrinol Metab, 2004. **89**(7): p. 3214-23.
26. Finley, D.J., B. Zhu, C.B. Barden, and T.J. Fahey, 3rd, *Discrimination of benign and malignant thyroid nodules by molecular profiling.* Ann Surg, 2004. **240**(3): p. 425-36; discussion 436-7.
27. Jarzab, B., M. Wiench, K. Fujarewicz, K. Simek, M. Jarzab, M. Oczko-Wojciechowska, J. Wloch, A. Czarniecka, E. Chmielik, D. Lange, A. Pawlaczek, S. Szpak, E. Gubala, and A. Swierniak, *Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications.* Cancer Res, 2005. **65**(4): p. 1587-97.
28. Arnaldi, L.A., R.C. Borra, R.M. Maciel, and J.M. Cerutti, *Gene Expression Profiles Reveal that DCN, DIO1, and DIO2 Are Underexpressed in Benign and Malignant Thyroid Tumours.* Thyroid, 2005. **15**(3): p. 210-21.

29. Chevillard, S., N. Ugolin, P. Vielh, K. Ory, C. Levalois, D. Elliott, G.L. Clayman, and A.K. El-Naggar, *Gene expression profiling of differentiated thyroid neoplasms: diagnostic and clinical implications.* Clin Cancer Res, 2004. **10**(19): p. 6586-97.
30. Yano, Y., N. Uematsu, T. Yashiro, H. Hara, E. Ueno, M. Miwa, G. Tsujimoto, Y. Aiyoshi, and K. Uchida, *Gene expression profiling identifies platelet-derived growth factor as a diagnostic molecular marker for papillary thyroid carcinoma.* Clin Cancer Res, 2004. **10**(6): p. 2035-43.
31. Strausberg, R.L., K.H. Buetow, S.F. Greenhut, L.H. Grouse, and C.F. Schaefer, *The cancer genome anatomy project: online resources to reveal the molecular signatures of cancer.* Cancer Invest, 2002. **20**(7-8): p. 1038-50.
32. Huang, Y., M. Prasad, W.J. Lemon, H. Hampel, F.A. Wright, K. Kornacker, V. LiVolsi, W. Frankel, R.T. Kloos, C. Eng, N.S. Pellegata, and A. de la Chapelle, *Gene expression in papillary thyroid carcinoma reveals highly consistent profiles.* Proc Natl Acad Sci U S A, 2001. **98**(26): p. 15044-9.
33. Aldred, M.A., Y. Huang, S. Liyanarachchi, N.S. Pellegata, O. Gimm, S. Jhiang, R.V. Davuluri, A. de la Chapelle, and C. Eng, *Papillary and follicular thyroid carcinomas show distinctly different microarray expression profiles and can be distinguished by a minimum of five genes.* J Clin Oncol, 2004. **22**(17): p. 3531-9.
34. Weber, F., L. Shen, M.A. Aldred, C.D. Morrison, A. Frilling, M. Saji, F. Schuppert, C.E. Broelsch, M.D. Ringel, and C. Eng, *Genetic classification of benign and malignant thyroid follicular neoplasia based on a 3-gene combination.* J Clin Endocrinol Metab, 2005. **90**(5): p. 2512-21.
35. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
36. Maere, S., K. Heymans, and M. Kuiper, *BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.* Bioinformatics, 2005. **21**(16): p. 3448-9.
37. Shannon, P., A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
38. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J R Stat Soc [Ser B], 1995. **57**(1): p. 289-300.
39. Lutzen, A., S.E. Liberti, and L.J. Rasmussen, *Cadmium inhibits human DNA mismatch repair in vivo.* Biochem Biophys Res Commun, 2004. **321**(1): p. 21-5.
40. Il'yasova, D. and G.G. Schwartz, *Cadmium and renal cancer.* Toxicol Appl Pharmacol, 2005. **207**(2): p. 179-86.
41. Sahmoun, A.E., L.D. Case, S.A. Jackson, and G.G. Schwartz, *Cadmium and prostate cancer: a critical epidemiologic analysis.* Cancer Invest, 2005. **23**(3): p. 256-63.
42. Chen, G.G., Z.M. Liu, A.C. Vlantis, G.M. Tse, B.C. Leung, and C.A. van Hasselt, *Heme oxygenase-1 protects against apoptosis induced by tumour necrosis factor-alpha and cycloheximide in papillary thyroid carcinoma cells.* J Cell Biochem, 2004. **92**(6): p. 1246-56.
43. Florianczyk, B., *Copper and metallothioneins in cancer cells.* Ann Univ Mariae Curie Sklodowska [Med], 2003. **58**(2): p. 390-3.
44. Nasulewicz, A., A. Mazur, and A. Opolski, *Role of copper in tumour angiogenesis--clinical implications.* J Trace Elem Med Biol, 2004. **18**(1): p. 1-8.

45. Kawanishi, S., Y. Hiraku, M. Murata, and S. Oikawa, *The role of metals in site-specific DNA damage with reference to carcinogenesis.* Free Radic Biol Med, 2002. **32**(9): p. 822-32.
46. Kucharzewski, M., J. Braziewicz, U. Majewska, and S. Gozdz, *Copper, zinc, and selenium in whole blood and thyroid tissue of people with various thyroid diseases.* Biol Trace Elem Res, 2003. **93**(1-3): p. 9-18.
47. Lotan, R., *Retinoids in cancer chemoprevention.* Faseb J, 1996. **10**(9): p. 1031-9.
48. Elisei, R., A. Vivaldi, L. Agate, R. Ciampi, E. Molinaro, P. Piampiani, C. Romei, P. Faviana, F. Basolo, P. Miccoli, A. Capodanno, P. Collecchi, F. Pacini, and A. Pinchera, *All-trans-retinoic acid treatment inhibits the growth of retinoic acid receptor beta messenger ribonucleic acid expressing thyroid cancer cell lines but does not reinduce the expression of thyroid-specific genes.* J Clin Endocrinol Metab, 2005. **90**(4): p. 2403-11.
49. Gruning, T., C. Tiepolt, K. Zophel, J. Bredow, J. Kropp, and W.G. Franke, *Retinoic acid for redifferentiation of thyroid cancer--does it hold its promise?* Eur J Endocrinol, 2003. **148**(4): p. 395-402.
50. Simon, D., J. Koehrle, C. Reiners, A.R. Boerner, C. Schmutzler, K. Mainz, P.E. Goretzki, and H.D. Roeher, *Redifferentiation therapy with retinoids: therapeutic option for advanced follicular and papillary thyroid carcinoma.* World J Surg, 1998. **22**(6): p. 569-74.
51. Specht, M.C., C.B. Barden, and T.J. Fahey, 3rd, *p44/p42-MAP kinase expression in papillary thyroid carcinomas.* Surgery, 2001. **130**(6): p. 936-40.
52. Maeta, H., S. Ohgi, and T. Terada, *Protein expression of matrix metalloproteinases 2 and 9 and tissue inhibitors of metalloproteinase 1 and 2 in papillary thyroid carcinomas.* Virchows Arch, 2001. **438**(2): p. 121-8.
53. Gonzalez-Sancho, J.M., V. Garcia, F. Bonilla, and A. Munoz, *Thyroid hormone receptors/THR genes in human cancer.* Cancer Lett, 2003. **192**(2): p. 121-32.
54. Belfiore, A., P. Gangemi, A. Costantino, G. Russo, G.M. Santonocito, O. Ippolito, M.F. Di Renzo, P. Comoglio, A. Fiumara, and R. Vigneri, *Negative/low expression of the Met/hepatocyte growth factor receptor identifies papillary thyroid carcinomas with high risk of distant metastases.* J Clin Endocrinol Metab, 1997. **82**(7): p. 2322-8.
55. Ippolito, A., V. Vella, G.L. La Rosa, G. Pellegriti, R. Vigneri, and A. Belfiore, *Immunostaining for Met/HGF receptor may be useful to identify malignancies in thyroid lesions classified suspicious at fine-needle aspiration biopsy.* Thyroid, 2001. **11**(8): p. 783-7.
56. Ramirez, R., D. Hsu, A. Patel, C. Fenton, C. Dinauer, R.M. Tuttle, and G.L. Francis, *Over-expression of hepatocyte growth factor/scatter factor (HGF/SF) and the HGF/SF receptor (cMET) are associated with a high risk of metastasis and recurrence for children and young adults with papillary thyroid carcinoma.* Clin Endocrinol (Oxf), 2000. **53**(5): p. 635-44.
57. Inaba, M., H. Sato, Y. Abe, S. Umemura, K. Ito, and H. Sakai, *Expression and significance of c-met protein in papillary thyroid carcinoma.* Tokai J Exp Clin Med, 2002. **27**(2): p. 43-9.
58. Di Renzo, M.F., M. Olivero, S. Ferro, M. Prat, I. Bongarzone, S. Pilotti, A. Belfiore, A. Costantino, R. Vigneri, M.A. Pierotti, and et al., *Overexpression of the c-MET/HGF receptor gene in human thyroid carcinomas.* Oncogene, 1992. **7**(12): p. 2549-53.
59. Hamada, A., S. Mankovskaya, V. Saenko, T. Rogounovitch, M. Mine, H. Namba, M. Nakashima, Y. Demidchik, E. Demidchik, and S. Yamashita, *Diagnostic usefulness of PCR profiling of the differentially expressed marker genes in thyroid papillary carcinomas.* Cancer Lett, 2005. **224**(2): p. 289-301.

60. Hawthorn, L., L. Stein, R. Varma, S. Wiseman, T. Loree, and D. Tan, *TIMP1 and SERPIN-A overexpression and TFF3 and CRABP1 underexpression as biomarkers for papillary thyroid carcinoma.* Head Neck, 2004. **26**(12): p. 1069-83.
61. Takano, T., A. Miyauchi, H. Yoshida, K. Kuma, and N. Amino, *High-throughput differential screening of mRNAs by serial analysis of gene expression: decreased expression of trefoil factor 3 mRNA in thyroid follicular carcinomas.* Br J Cancer, 2004. **90**(8): p. 1600-5.
62. Takano, T., A. Miyauchi, H. Yoshida, K. Kuma, and N. Amino, *Decreased relative expression level of trefoil factor 3 mRNA to galectin-3 mRNA distinguishes thyroid follicular carcinoma from adenoma.* Cancer Lett, 2005. **219**(1): p. 91-6.
63. Poblete, M.T., F. Nualart, M. del Pozo, J.A. Perez, and C.D. Figueroa, *Alpha 1-antitrypsin expression in human thyroid papillary carcinoma.* Am J Surg Pathol, 1996. **20**(8): p. 956-63.
64. Wasenius, V.M., S. Hemmer, E. Kettunen, S. Knuutila, K. Franssila, and H. Joensuu, *Hepatocyte growth factor receptor, matrix metalloproteinase-11, tissue inhibitor of metalloproteinase-1, and fibronectin are up-regulated in papillary thyroid carcinoma: a cDNA and tissue microarray study.* Clin Cancer Res, 2003. **9**(1): p. 68-75.
65. Hesse, E., P.B. Musholt, E. Potter, T. Petrich, M. Wehmeier, R. von Wasielewski, R. Lichtinghagen, and T.J. Musholt, *Oncofoetal fibronectin--a tumour-specific marker in detecting minimal residual disease in differentiated thyroid carcinoma.* Br J Cancer, 2005. **93**(5): p. 565-70.
66. Liu, W., S.L. Asa, and S. Ezzat, *1alpha,25-Dihydroxyvitamin D3 targets PTEN-dependent fibronectin expression to restore thyroid cancer cell adhesiveness.* Mol Endocrinol, 2005. **19**(9): p. 2349-57.
67. Prasad, M.L., N.S. Pellegata, Y. Huang, H.N. Nagaraja, A. de la Chapelle, and R.T. Kloos, *Galectin-3, fibronectin-1, CITED-1, HBME1 and cytokeratin-19 immunohistochemistry is useful for the differential diagnosis of thyroid tumours.* Mod Pathol, 2005. **18**(1): p. 48-57.
68. Lazar, V., J.M. Bidart, B. Caillou, C. Mahe, L. Lacroix, S. Filetti, and M. Schlumberger, *Expression of the Na+/I- symporter gene in human thyroid tumours: a comparison study with other thyroid-specific genes.* J Clin Endocrinol Metab, 1999. **84**(9): p. 3228-34.
69. Segev, D.L., D.P. Clark, M.A. Zeiger, and C. Umbricht, *Beyond the suspicious thyroid fine needle aspirate. A review.* Acta Cytol, 2003. **47**(5): p. 709-22.
70. Bergstrom, J.D., B. Westermark, and N.E. Heldin, *Epidermal growth factor receptor signalling activates met in human anaplastic thyroid carcinoma cells.* Exp Cell Res, 2000. **259**(1): p. 293-9.
71. Huang, Y., A. de la Chapelle, and N.S. Pellegata, *Hypermethylation, but not LOH, is associated with the low expression of MT1G and CRABP1 in papillary thyroid carcinoma.* Int J Cancer, 2003. **104**(6): p. 735-44.
72. O'Donovan, N., A. Fischer, E.M. Abdo, F. Simon, H.J. Peter, H. Gerber, U. Buergi, and U. Marti, *Differential expression of IgG Fc binding protein (FcgammaBP) in human normal thyroid tissue, thyroid adenomas and thyroid carcinomas.* J Endocrinol, 2002. **174**(3): p. 517-24.
73. Barden, C.B., K.W. Shister, B. Zhu, G. Guiter, D.Y. Greenblatt, M.A. Zeiger, and T.J. Fahey, 3rd, *Classification of follicular thyroid tumours by molecular signature: results of gene profiling.* Clin Cancer Res, 2003. **9**(5): p. 1792-800.
74. Finn, S.P., P. Smyth, S. Cahill, C. Streck, E.M. O'Regan, R. Flavin, J. Sherlock, D. Howells, R. Henfrey, M. Cullen, M. Toner, C. Timon, J.J. O'Leary, and O.M. Sheils, *Expression microarray analysis of papillary thyroid carcinoma and benign thyroid*

tissue: emphasis on the follicular variant and potential markers of malignancy. Virchows Arch, 2007. **450**(3): p. 249-60.
75. Lubitz, C.C., L.A. Gallagher, D.J. Finley, B. Zhu, and T.J. Fahey, 3rd, *Molecular analysis of minimally invasive follicular carcinomas by gene profiling.* Surgery, 2005. **138**(6): p. 1042-8; discussion 1048-9.
76. Lubitz, C.C., S.K. Ugras, J.J. Kazam, B. Zhu, T. Scognamiglio, Y.T. Chen, and T.J. Fahey, 3rd, *Microarray analysis of thyroid nodule fine-needle aspirates accurately classifies benign and malignant lesions.* J Mol Diagn, 2006. **8**(4): p. 490-8; quiz 528.
77. Eszlinger, M., K. Krohn, A. Kukulska, B. Jarzab, and R. Paschke, *Perspectives and limitations of microarray-based gene expression profiling of thyroid tumours.* Endocr Rev, 2007. **28**(3): p. 322-38.
78. Fujarewicz, K., M. Jarzab, M. Eszlinger, K. Krohn, R. Paschke, M. Oczko-Wojciechowska, M. Wiench, A. Kukulska, B. Jarzab, and A. Swierniak, *A multi-gene approach to differentiate papillary thyroid carcinoma from benign lesions: gene selection using support vector machines with bootstrapping.* Endocr Relat Cancer, 2007. **14**(3): p. 809-26.
79. Eszlinger, M., K. Krohn, R. Frenzel, S. Kropf, A. Tonjes, and R. Paschke, *Gene expression analysis reveals evidence for inactivation of the TGF-beta signalling cascade in autonomously functioning thyroid nodules.* Oncogene, 2004. **23**(3): p. 795-804.
80. Eszlinger, M., K. Krohn, and R. Paschke, *Complementary DNA expression array analysis suggests a lower expression of signal transduction proteins and receptors in cold and hot thyroid nodules.* J Clin Endocrinol Metab, 2001. **86**(10): p. 4834-42.
81. Rhodes, D.R., S. Kalyana-Sundaram, V. Mahavisno, R. Varambally, J. Yu, B.B. Briggs, T.R. Barrette, M.J. Anstet, C. Kincead-Beal, P. Kulkarni, S. Varambally, D. Ghosh, and A.M. Chinnaiyan, *Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles.* Neoplasia, 2007. **9**(2): p. 166-80.
82. Lin, J.D. and T.C. Chao, *Vascular endothelial growth factor in thyroid cancers.* Cancer Biother Radiopharm, 2005. **20**(6): p. 648-61.
83. Siemann, D.W., D.J. Chaplin, and M.R. Horsman, *Vascular-targeting therapies for treatment of malignant disease.* Cancer, 2004. **100**(12): p. 2491-9.
84. Huang, S.M., J.C. Lee, T.J. Wu, and N.H. Chow, *Clinical relevance of vascular endothelial growth factor for thyroid neoplasms.* World J Surg, 2001. **25**(3): p. 302-6.
85. Klein, M., E. Picard, J.M. Vignaud, B. Marie, L. Bresler, B. Toussaint, G. Weryha, A. Duprez, and J. Leclere, *Vascular endothelial growth factor gene and protein: strong expression in thyroiditis and thyroid carcinoma.* J Endocrinol, 1999. **161**(1): p. 41-9.
86. Viglietto, G., D. Maglione, M. Rambaldi, J. Cerutti, A. Romano, F. Trapasso, M. Fedele, P. Ippolito, G. Chiappetta, G. Botti, and et al., *Upregulation of vascular endothelial growth factor (VEGF) and downregulation of placenta growth factor (PlGF) associated with malignancy in human thyroid tumours and cell lines.* Oncogene, 1995. **11**(8): p. 1569-79.
87. Liu, F.T. and G.A. Rabinovich, *Galectins as modulators of tumour progression.* Nat Rev Cancer, 2005. **5**(1): p. 29-41.
88. Collet, J.F., I. Hurbain, C. Prengel, O. Utzmann, F. Scetbon, J.F. Bernaudin, and A. Fajac, *Galectin-3 immunodetection in follicular thyroid neoplasms: a prospective study on fine-needle aspiration samples.* Br J Cancer, 2005. **93**(10): p. 1175-81.
89. Bartolazzi, A., A. Gasbarri, M. Papotti, G. Bussolati, T. Lucante, A. Khan, H. Inohara, F. Marandino, F. Orlandi, F. Nardi, A. Vecchione, R. Tecce, and O. Larsson, *Application of an immunodiagnostic method for improving preoperative diagnosis of nodular thyroid lesions.* Lancet, 2001. **357**(9269): p. 1644-50.

90. Shin, E., W.Y. Chung, W.I. Yang, C.S. Park, and S.W. Hong, *RET/PTC and CK19 expression in papillary thyroid carcinoma and its clinicopathologic correlation.* J Korean Med Sci, 2005. **20**(1): p. 98-104.
91. Rorive, S., B. Eddafali, S. Fernandez, C. Decaestecker, S. Andre, H. Kaltner, I. Kuwabara, F.T. Liu, H.J. Gabius, R. Kiss, and I. Salmon, *Changes in galectin-7 and cytokeratin-19 expression during the progression of malignancy in thyroid tumours: diagnostic and biological implications.* Mod Pathol, 2002. **15**(12): p. 1294-301.
92. Choi, Y.L., M.K. Kim, J.W. Suh, J. Han, J.H. Kim, J.H. Yang, and S.J. Nam, *Immunoexpression of HBME-1, high molecular weight cytokeratin, cytokeratin 19, thyroid transcription factor-1, and E-cadherin in thyroid carcinomas.* J Korean Med Sci, 2005. **20**(5): p. 853-9.
93. Scognamiglio, T., E. Hyjek, J. Kao, and Y.T. Chen, *Diagnostic usefulness of HBME1, galectin-3, CK19, and CITED1 and evaluation of their expression in encapsulated lesions with questionable features of papillary thyroid carcinoma.* Am J Clin Pathol, 2006. **126**(5): p. 700-8.
94. Heinlein, C.A. and C. Chang, *Androgen receptor in prostate cancer.* Endocr Rev, 2004. **25**(2): p. 276-308.
95. Blechet, C., P. Lecomte, L. De Calan, P. Beutter, and S. Guyetant, *Expression of sex steroid hormone receptors in C cell hyperplasia and medullary thyroid carcinoma.* Virchows Arch, 2007. **450**(4): p. 433-9.
96. Kimura, M., S. Kotani, T. Hattori, N. Sumi, T. Yoshioka, K. Todokoro, and Y. Okano, *Cell cycle-dependent expression and spindle pole localization of a novel human protein kinase, Aik, related to Aurora of Drosophila and yeast Ipl1.* J Biol Chem, 1997. **272**(21): p. 13766-71.
97. Zhou, H., J. Kuang, L. Zhong, W.L. Kuo, J.W. Gray, A. Sahin, B.R. Brinkley, and S. Sen, *Tumour amplified kinase STK15/BTAK induces centrosome amplification, aneuploidy and transformation.* Nat Genet, 1998. **20**(2): p. 189-93.
98. Wiseman, S.M., H. Masoudi, P. Niblock, D. Turbin, A. Rajput, J. Hay, S. Bugis, D. Filipenko, D. Huntsman, and B. Gilks, *Anaplastic thyroid carcinoma: expression profile of targets for therapy offers new insights for disease treatment.* Ann Surg Oncol, 2007. **14**(2): p. 719-29.
99. Rosen, J., M. He, C. Umbricht, H.R. Alexander, A.P. Dackiw, M.A. Zeiger, and S.K. Libutti, *A six-gene model for differentiating benign from malignant thyroid tumours on the basis of gene expression.* Surgery, 2005. **138**(6): p. 1050-6; discussion 1056-7.
100. Rossi, E.D., M. Raffaelli, A. Mule, A. Miraglia, C.P. Lombardi, F.M. Vecchio, and G. Fadda, *Simultaneous immunohistochemical expression of HBME-1 and galectin-3 differentiates papillary carcinomas from hyperfunctioning lesions of the thyroid.* Histopathology, 2006. **48**(7): p. 795-800.
101. Kebebew, E., M. Peng, E. Reiff, and A. McMillan, *Diagnostic and extent of disease multigene assay for malignant thyroid neoplasms.* Cancer, 2006. **106**(12): p. 2592-7.
102. Barroeta, J.E., Z.W. Baloch, P. Lal, T.L. Pasha, P.J. Zhang, and V.A. LiVolsi, *Diagnostic value of differential expression of CK19, Galectin-3, HBME-1, ERK, RET, and p16 in benign and malignant follicular-derived lesions of the thyroid: an immunohistochemical tissue microarray analysis.* Endocr Pathol, 2006. **17**(3): p. 225-34.
103. Nakamura, N., L.A. Erickson, L. Jin, S. Kajita, H. Zhang, X. Qian, K. Rumilla, and R.V. Lloyd, *Immunohistochemical separation of follicular variant of papillary thyroid carcinoma from follicular adenoma.* Endocr Pathol, 2006. **17**(3): p. 213-23.

104. Melck, A.L., H. Masoudi, O.L. Griffith, A. Rajput, G.E. Wilkins, S. Bugis, S. Jones, and S.M. Wiseman, *Cell Cycle Regulators Show Diagnostic and Prognostic Utility for Differentiated Thyroid Cancer.* Ann Surg Oncol, 2007. **14**(12): p. 3403–11.
105. Cerutti, J.M., F.R. Latini, C. Nakabashi, R. Delcelo, V.P. Andrade, M.J. Amadei, R.M. Maciel, F.C. Hojaij, D. Hollis, J. Shoemaker, and G.J. Riggins, *Diagnosis of suspicious thyroid nodules using four protein biomarkers.* Clin Cancer Res, 2006. **12**(11 Pt 1): p. 3311-8.
106. Smith, N.D., J.N. Rubenstein, S.E. Eggener, and J.M. Kozlowski, *The p53 tumour suppressor gene and nuclear protein: basic science review and relevance in the management of bladder cancer.* J Urol, 2003. **169**(4): p. 1219-28.
107. Bargonetti, J. and J.J. Manfredi, *Multiple roles of the tumour suppressor p53.* Curr Opin Oncol, 2002. **14**(1): p. 86-91.
108. Blagosklonny, M.V., P. Giannakakou, M. Wojtowicz, L.Y. Romanova, K.B. Ain, S.E. Bates, and T. Fojo, *Effects of p53-expressing adenovirus on the chemosensitivity and differentiation of anaplastic thyroid cancer cells.* J Clin Endocrinol Metab, 1998. **83**(7): p. 2516-22.
109. Lowe, S.W., *Cancer therapy and p53.* Curr Opin Oncol, 1995. **7**(6): p. 547-53.
110. Lam, K.Y., C.Y. Lo, K.W. Chan, and K.Y. Wan, *Insular and anaplastic carcinoma of the thyroid: a 45-year comparative study at a single institution and a review of the significance of p53 and p21.* Ann Surg, 2000. **231**(3): p. 329-38.
111. Urruticoechea, A., I.E. Smith, and M. Dowsett, *Proliferation marker Ki-67 in early breast cancer.* J Clin Oncol, 2005. **23**(28): p. 7212-20.
112. Kjellman, P., G. Wallin, A. Hoog, G. Auer, C. Larsson, and J. Zedenius, *MIB-1 index in thyroid tumours: a predictor of the clinical course in papillary thyroid carcinoma.* Thyroid, 2003. **13**(4): p. 371-80.
113. Karayan-Tapon, L., E. Menet, J. Guilhot, P. Levillain, C.J. Larsen, and J.L. Kraimps, *Topoisomerase II alpha and telomerase expression in papillary thyroid carcinomas.* Eur J Surg Oncol, 2004. **30**(1): p. 73-9.
114. Lee, A., V.A. LiVolsi, and Z.W. Baloch, *Expression of DNA topoisomerase IIalpha in thyroid neoplasia.* Mod Pathol, 2000. **13**(4): p. 396-400.
115. Cardoso, F., V. Durbecq, D. Larsimont, M. Paesmans, J.Y. Leroy, G. Rouas, C. Sotiriou, N. Renard, V. Richard, M.J. Piccart, and A. Di Leo, *Correlation between complete response to anthracycline-based chemotherapy and topoisomerase II-alpha gene amplification and protein overexpression in locally advanced/metastatic breast cancer.* Int J Oncol, 2004. **24**(1): p. 201-9.
116. Jarvinen, T.A. and E.T. Liu, *Topoisomerase IIalpha gene (TOP2A) amplification and deletion in cancer--more common than anticipated.* Cytopathology, 2003. **14**(6): p. 309-13.
117. Ain, K.B., *Anaplastic thyroid carcinoma: a therapeutic challenge.* Semin Surg Oncol, 1999. **16**(1): p. 64-9.
118. Are, C. and A.R. Shaha, *Anaplastic thyroid carcinoma: biology, pathogenesis, prognostic factors, and treatment approaches.* Ann Surg Oncol, 2006. **13**(4): p. 453-64.
119. O'Neill, J.P., B. O'Neill, C. Condron, M. Walsh, and D. Bouchier-Hayes, *Anaplastic (undifferentiated) thyroid cancer: improved insight and therapeutic strategy into a highly aggressive disease.* J Laryngol Otol, 2005. **119**(8): p. 585-91.
120. Fluge, O., O. Bruland, L.A. Akslen, J.R. Lillehaug, and J.E. Varhaug, *Gene expression in poorly differentiated papillary thyroid carcinomas.* Thyroid, 2006. **16**(2): p. 161-75.

121. Saltman, B., B. Singh, C.V. Hedvat, V.B. Wreesmann, and R. Ghossein, *Patterns of expression of cell cycle/apoptosis genes along the spectrum of thyroid carcinoma progression.* Surgery, 2006. **140**(6): p. 899-905; discussion 905-6.
122. Ordonez, N.G., A.K. El-Naggar, R.C. Hickey, and N.A. Samaan, *Anaplastic thyroid carcinoma. Immunocytochemical study of 32 cases.* Am J Clin Pathol, 1991. **96**(1): p. 15-24.
123. Van Aken, E., O. De Wever, A.S. Correia da Rocha, and M. Mareel, *Defective E-cadherin/catenin complexes in human cancer.* Virchows Arch, 2001. **439**(6): p. 725-51.
124. Ilyas, M., *Wnt signalling and the mechanistic basis of tumour development.* J Pathol, 2005. **205**(2): p. 130-44.
125. Wiseman, S.M., H. Masoudi, P. Niblock, D. Turbin, A. Rajput, J. Hay, D. Filipenko, D. Huntsman, and B. Gilks, *Derangement of the E-cadherin/catenin complex is involved in transformation of differentiated to anaplastic thyroid carcinoma.* Am J Surg, 2006. **191**(5): p. 581-7.
126. Pollina, L., F. Pacini, G. Fontanini, S. Vignati, G. Bevilacqua, and F. Basolo, *bcl-2, p53 and proliferating cell nuclear antigen expression is related to the degree of differentiation in thyroid carcinomas.* Br J Cancer, 1996. **73**(2): p. 139-43.
127. Pilotti, S., P. Collini, F. Rilke, G. Cattoretti, R. Del Bo, and M.A. Pierotti, *Bcl-2 protein expression in carcinomas originating from the follicular epithelium of the thyroid gland.* J Pathol, 1994. **172**(4): p. 337-42.
128. Kim, R., K. Tanabe, Y. Uchida, M. Emi, and T. Toge, *Effect of Bcl-2 antisense oligonucleotide on drug-sensitivity in association with apoptosis in undifferentiated thyroid carcinoma.* Int J Mol Med, 2003. **11**(6): p. 799-804.
129. Vieira, J.M., S.C. Santos, C. Espadinha, I. Correia, T. Vag, C. Casalou, B.M. Cavaco, A.L. Catarino, S. Dias, and V. Leite, *Expression of vascular endothelial growth factor (VEGF) and its receptors in thyroid carcinomas of follicular origin: a potential autocrine loop.* Eur J Endocrinol, 2005. **153**(5): p. 701-9.
130. Bauer, A.J., R. Terrell, N.K. Doniparthi, A. Patel, R.M. Tuttle, M. Saji, M.D. Ringel, and G.L. Francis, *Vascular endothelial growth factor monoclonal antibody inhibits growth of anaplastic thyroid cancer xenografts in nude mice.* Thyroid, 2002. **12**(11): p. 953-61.
131. Dziba, J.M., R. Marcinek, G. Venkataraman, J.A. Robinson, and K.B. Ain, *Combretastatin A4 phosphate has primary antineoplastic activity against human anaplastic thyroid carcinoma cell lines and xenograft tumours.* Thyroid, 2002. **12**(12): p. 1063-70.
132. Straight, A.M., K. Oakley, R. Moores, A.J. Bauer, A. Patel, R.M. Tuttle, J. Jimeno, and G.L. Francis, *Aplidin reduces growth of anaplastic thyroid cancer xenografts and the expression of several angiogenic genes.* Cancer Chemother Pharmacol, 2006. **57**(1): p. 7-14.
133. Shaha, A., *Treatment of thyroid cancer based on risk groups.* J Surg Oncol, 2006. **94**(8): p. 683-91.
134. Mazzanti, C., M.A. Zeiger, N.G. Costouros, C. Umbricht, W.H. Westra, D. Smith, H. Somervell, G. Bevilacqua, H.R. Alexander, and S.K. Libutti, *Using gene expression profiling to differentiate benign versus malignant thyroid tumours.* Cancer Res, 2004. **64**(8): p. 2898-903.
135. Cerutti, J.M., R. Delcelo, M.J. Amadei, C. Nakabashi, R.M. Maciel, B. Peterson, J. Shoemaker, and G.J. Riggins, *A preoperative diagnostic test that distinguishes benign from malignant thyroid carcinoma based on gene expression.* J Clin Invest, 2004. **113**(8): p. 1234-42.

136. Pauws, E., G.J. Veenboer, J.W. Smit, J.J. de Vijlder, H. Morreau, and C. Ris-Stalpers, *Genes differentially expressed in thyroid carcinoma identified by comparison of SAGE expression profiles.* Faseb J, 2004. **18**(3): p. 560-1.
137. Zou, M., K.S. Famulski, R.S. Parhar, E. Baitei, F.A. Al-Mohanna, N.R. Farid, and Y. Shi, *Microarray analysis of metastasis-associated gene expression profiling in a murine model of thyroid carcinoma pulmonary metastasis: identification of S100A4 (Mts1) gene overexpression as a poor prognostic marker for thyroid carcinoma.* J Clin Endocrinol Metab, 2004. **89**(12): p. 6146-54.
138. Chen, K.T., J.D. Lin, T.C. Chao, C. Hsueh, C.A. Chang, H.F. Weng, and E.C. Chan, *Identifying differentially expressed genes associated with metastasis of follicular thyroid cancer by cDNA expression array.* Thyroid, 2001. **11**(1): p. 41-6.
139. Scarpino, S., F. Cancellario d'Alena, A. Di Napoli, A. Pasquini, A. Marzullo, and L.P. Ruco, *Increased expression of Met protein is associated with up-regulation of hypoxia inducible factor-1 (HIF-1) in tumour cells in papillary carcinoma of the thyroid.* J Pathol, 2004. **202**(3): p. 352-8.
140. Nardone, H.C., A.F. Ziober, V.A. LiVolsi, S.J. Mandel, Z.W. Baloch, R.S. Weber, R. Mick, and B.L. Ziober, *c-Met expression in tall cell variant papillary carcinoma of the thyroid.* Cancer, 2003. **98**(7): p. 1386-93.
141. Mineo, R., A. Costantino, F. Frasca, L. Sciacca, S. Russo, R. Vigneri, and A. Belfiore, *Activation of the hepatocyte growth factor (HGF)-Met system in papillary thyroid cancer: biological effects of HGF in thyroid cancer cells depend on Met expression levels.* Endocrinology, 2004. **145**(9): p. 4355-65.
142. Shi, Y., R.S. Parhar, M. Zou, M.M. Hammami, M. Akhtar, Z.P. Lum, N.R. Farid, S.T. Al-Sedairy, and M.C. Paterson, *Tissue inhibitor of metalloproteinases-1 (TIMP-1) mRNA is elevated in advanced stages of thyroid carcinoma.* Br J Cancer, 1999. **79**(7-8): p. 1234-9.
143. Umeki, K., T. Tanaka, I. Yamamoto, Y. Aratake, T. Kotani, F. Sakamoto, S. Noguchi, and S. Ohtaki, *Differential expression of dipeptidyl peptidase IV (CD26) and thyroid peroxidase in neoplastic thyroid tissues.* Endocr J, 1996. **43**(1): p. 53-60.
144. Chan, S.K., O.L. Griffith, I.T. Tai, and S.J.M. Jones, *Meta-analysis of Colorectal Cancer Gene Expression Profiling Studies Identifies Consistently Reported Candidate Biomarkers.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(3): p. 543–52**.**

5. ORegAnno: an open-access community-driven resource for regulatory annotation[11,12,13]

5.1. Introduction

A consequence of the escalating pace of genomic sequencing has been the requirement for novel methodology and large-scale efforts to interpret and annotate sequence function. Initial efforts to achieve this were primarily focused on identifying protein-coding genes, RNA genes and repetitive DNA, since the rules governing their presence are generally tractable. However, less annotated, due to their small size and variability, gene regulatory sequences are widely regarded to be at least as important to our understanding of biological systems. To aid in their identification, computational techniques such as phylogenetic footprinting, transcription factor (TF) binding matrices, and motif clustering have been developed [1-3]. Unfortunately, the predictive ability of such methods has been difficult to assess without large, well-described, and comprehensive collections of biologically validated regulatory sequences [3]. Sets of *cis*-regulatory sequences have been annotated by curation from the primary literature and several databases have been developed to collect and disseminate these sets [4-11]. However, these databases are often species- or process-specific, do not provide sufficient details about the experiments or conditions under which function was demonstrated, and in some cases require payment for access. Data access is generally limited to web-based search pages without any option for the programmatic interaction essential to most bioinformatics studies. Finally, they are typically 'closed systems' in that they do not allow continued addition or annotation by the research community and as such are not maintainable over the long-term without vast resources. We have developed the Open Regulatory Annotation database (ORegAnno) to overcome these

[11] A portion of this chapter has been published. Griffith OL*, Montgomery SB*, Bernier B, Chu B, Aerts S, Sleumer MC, Bilenky M, Haeussler M, Griffith M, Gallo SM, Giardine B, Mahony S, Hooghe B, Van Loo P, Blanco E, Ticoll A, Lithwick S, Portales-Casamar E, Donaldson IJ, Robertson G, Wadelius C, De Bleser P, Vlieghe D, Halfon MS, Wasserman W, Hardison R, Bergman CM, Jones SJM; The Open Regulatory Annotation Consortium. 2008. ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res. 36(Database issue):D107-13. *These authors contributed equally to this work.

[12] A portion of this chapter has been published. Montgomery SB*, Griffith OL*, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJM. 2006. ORegAnno: An open access database and curation system for literature-derived promoters, transcription factor binding sites, and regulatory variation. Bioinformatics. 22(5):637-40. *These authors contributed equally to this work.

[13] Co-authorship details: I worked together with Stephen Montgomery to create the ORegAnno resource. Stephen Montgomery was primarily responsible for code development. I was primarily responsible for organization of data entry and creation of the evidence ontology. Stephen Montgomery wrote the majority of the original paper (Montgomery *et al*., 2006) and I wrote the majority of the second follow-up paper (Griffith *et al*., 2008). However, we are listed as joint first-authors on both because overall our contributions were approximately equal for each. I was responsible for all text, figures and tables included in this chapter except where indicated below. Stephen Montgomery contributed to sections 5.1, 5.2.1, 5.2.2 and 5.5. Stein Aerts and Casey Bergman contributed to section 5.2.5 and were primarily responsible for the separate study described therein. Many others (see author lists on the publications) contributed data, functionality, editorial suggestions, supervision, or funding.

challenges and support research in regulatory biology [12]. ORegAnno provides standardized technologies for the long-term, community-driven, open-access curation of *cis*-regulatory data.

5.2. Description of the ORegAnno database

5.2.1. Database overview

ORegAnno (http://www.oreganno.org) is a database and literature curation system for community-based annotation of experimentally proven DNA regulatory regions, transcription factor binding sites (TFBS), and regulatory variants. A 'publication queue' allows papers of interest to be added to the system for future curation. Thus both regulatory papers and their regulatory sequences are managed in the system. ORegAnno is based on open-source technology and is comprised of a MySQL database (Figure 5.1) with a Java-based web application that indexes new annotations using the Lucene search engine (http://lucene.apache.org/) and provides programmatic access to the underlying data using Hibernate (http://www.hibernate.org/) and SOAP Web Services.

In Figure 5.2 an outline of the annotation process and information flow is presented. Users in the gene regulation community can enter or 'check out' papers from the publication queue for detailed manual curation using a series of annotation pages. A typical record entry consists of species, sequence type, sequence (plus sufficient flanking sequence for genome alignment), target gene, binding factor, experimental outcome and one or more detailed lines of experimental evidence demonstrating function of the sequence (see section 5.2.2 for more details). Records are cross-referenced to Ensembl [13] or Entrez Gene identifiers [14], PubMed [15], and dbSNP [16] (for regulatory polymorphisms). Before committing a record to the database, ORegAnno performs a number of error checks (e.g., that the sequence has not been entered previously) and asks the user to verify its contents before submission. Once submitted, the record is added to the database and an email is generated containing an XML representation of this record to members of the ORegAnno developers' mailing-list (oreganno-guts@bcgsc.ca). A BLAST-based mapping agent then assigns genome coordinates to each sequence, allowing it to be viewed as a track in the Ensembl or UCSC genome browsers. Once finished with a paper, a user will then 'close' it in the queue and assign an annotation result (success, neutral, or failure). Existing records can be updated, commented, and scored (positive if verified as correct; negative if a problem is identified) by any registered user or deprecated and replaced by a "validator" user (explained

further below). The complete database or any subset can be searched or downloaded in a number of formats or accessed programmatically.

ORegAnno permits open annotation of regulatory regions by providing roles and secure user accounts to contributors. Three roles exist for ORegAnno contributors: 'User', 'validator' and 'administrator'. A user role enables a contributor to add individual annotations to the database. A validator role enables a contributor to score individual annotations in the database. Validators will modify an overall score for an annotation based on their ability to confirm the reliability of annotation from literature. Validators have the option of increasing the annotation score by one if they can confirm the record, leaving the score unchanged if their conclusions are indeterminate, or decreasing the score by one if an error has been found. Each observation and score modification of an annotation along with the associated validator user information is stored in ORegAnno. An Administrator role enables a contributor to assign roles, add or define new evidence (classes, types, and subtypes) and batch upload large sets of annotations directly to the database. Both administrator and validator roles allow the modification of records; for a record modification, a new record is created and the old record is marked as being deprecated by the newer record. Each role is further permitted to add comments to individual annotations to improve subsequent users' understanding of a particular annotation. ORegAnno's usage of roles provides a level of accountability in the database as users become owners of their annotation and validators become responsible for verifying an annotation's authenticity.

ORegAnno comes equipped with analysis tools to assist in annotation of new records. In many cases, extracting genome sequence from literature and identifying the corresponding sequences in genome databases is problematic [17]. ORegAnno provides the tools ENSSCAN for finding one or more specific sequences within distances relative to the start of an EnsEMBL transcript, ENSFETCH for retrieving small sequences within distances relative to the start of an EnsEMBL transcript (i.e. from -34 to -40 of the transcription start site), NCBISCAN for finding one or more specific sequences within defined distances of a GenBank-reference sequence, and NCBIFETCH for highlighting small (gapped) sequences within a GenBank-reference sequence.

The ORegAnno website has been recently updated to use Ajax technology, improving the ease of annotation. Ajax improves a web page's usability by exchanging small amounts of data with the server behind the scenes, so that the entire web page does not have to be reloaded each time the

user requests a change (http://www.xul.fr/en-xml-ajax.html). Detailed case studies and documentation are available through the help pages to guide users through the entire process of annotating a paper. ORegAnno web pages also feature individual 'help bubbles' to explain each field in the record and queue annotation processes.

5.2.2. Data model for an ORegAnno record

For each type of annotation that is currently in ORegAnno, the database obeys the following rules:

1) Each annotation describes a regulatory property of one target gene which is either user-defined, in Entrez Gene or in EnsEMBL.
2) Each annotation must be attributed to a species which has a taxonomy id in the NCBI Taxonomy database (e.g., 9606 for Homo sapiens).
3) Each annotation can optionally be associated to a specific dataset (e.g., ChIPSeq STAT1 experiment-derived sites). This functionality allows external curators to manage particular sets of annotation using ORegAnno's curation tools.
4) Each annotation specifies an evidence type, subtype and class describing the biological technique cited to discover the regulatory sequence. Each piece of experimental evidence is also optionally associated to a specific cell type using the eVOC cell type ontology [18].
5) Each transcription factor binding site or regulatory mutation must specify a target transcription factor which is either user-defined, in Entrez Gene or in EnsEMBL. If there is no recorded gene target, a classification of "unknown" is specified.
6) Each transcription factor binding site or regulatory mutation must include sequence with at least 40 bases of flanking genomic sequence (generally much more than this) to allow the site to be mapped to any release of an associated genome.
7) Where available, any annotation can provide search space information specifying the sequence region that was assayed, not just the functional regulatory sequence itself.
8) User information is recorded with each annotation.
9) Each annotation must reference a valid PubMed article. To reduce the entry of redundant annotations, a warning is issued if an annotation is found with either an existing reference identifier or matching genomic sequence. The PubMed article must also exist in the publication queue before it can be annotated.
10) For regulatory mutations, each variant that has been proven to cause a change in gene expression is a separate record. The sequences containing both the wild-type and mutant

sequences must be specified. If available, a dbSNP cross-reference can also be specified. The type of variant is specified as either being germline, somatic or artificial.

11) Each record is associated to a positive, neutral or negative outcome based on the experimental results from the primary reference. For instance, a sequence that was demonstrated not to bind a particular transcription factor could be annotated as a negative outcome; however, to be meaningful, the associated evidence must provide adequate information to determine the conditions assayed.

5.2.3. Ontologies in ORegAnno

The ORegAnno evidence ontology is a simple ontology of evidence classes, types and subtypes for describing experiments that demonstrate the identity and/or function of regulatory sequences and their factors. These lines of evidence capture critical details from primary experiments and allow end users to filter the ORegAnno sequence set based on their own criteria for experimental credibility. Evidence classes are broken into two categories: the 'regulator' classes describe evidence for the specific protein(s) that bind a site. The 'regulatory site' classes describe evidence for the function of a regulatory sequence itself. These two categories are further divided into three levels of regulation (transcription, transcript stability and translation). ORegAnno is primarily a database of experimentally verified regulatory sequences. Therefore, there should always be at least one piece of evidence with the regulatory site class. The evidence for a site might involve mutating the putative site and measuring the effect of the mutation. In addition, the user can enter evidence for the identity of the regulator (e.g. a transcription factor) that actually binds the regulatory site. A regulator site is not required to have evidence of a specific regulator protein. In fact, in some cases the site may function by mechanisms not requiring a protein (DNA-RNA interactions, DNA-DNA interactions, etc). Or, the regulator may not have been determined yet. Evidence types describe the generic assay used while subtypes define specific implementations of these assays. Each annotation can have multiple entries from any evidence class, type and subtype describing each piece of experimental evidence for the regulatory sequence and/or binding protein. The ontology has been considerably developed and extended since first published and currently consists of 6 classes (Table 5.1), 14 evidence types and 70 evidence subtypes (Table 5.2). This ontology has been adopted by the PAZAR resource and is being developed in collaboration with that group using Protégé (http://protege.stanford.edu/). PAZAR is an open-source/open-access 'boutique system' for sharing of regulatory sequence annotation collections such as OregAnno [19]. The complete

evidence ontology can be obtained in XML format (http://www.oreganno.org/oregano/evidence.xml) or as a Protégé project file (http://www.pazar.info/ontologies/newevidence.pprj). Within each line of evidence, a user can also specify the cell type in which experiments were conducted using the eVOC cell type ontology [18]. We are working to incorporate additional Ontologies such as the BRENDA Tissue Ontology, and improvements to the Sequence Ontology are currently being developed for the *cis*-regulatory domain.

5.2.4. Publication queue

An important feature of ORegAnno called the 'publication queue' was created as a literature management system to allow registered users to input relevant papers from the scientific literature as targets for annotation. All that is required to enter a publication is a valid PubMed identifier. Optionally, a TF can be specified, allowing users to later search the queue for papers related to TFs of interest. Normally, publications are added to the queue with an entry type of 'expert entry', indicating that a human expert reviewed the paper and found it to be relevant. However, it is also possible to enter 'text-mining entry' papers (see below). A publication enters the queue with an initial state of 'pending'. Any user can then 'open' the publication and begin the annotation process. Once annotated, the paper is either 'closed' or reset to 'pending' if annotation work remains. Free-form comment fields are optional for each change of state. However, when a publication is closed, one of several standardized closure comments must be chosen. The four possible closure comments are: "Success - addition of new records", "Failure - did not describe regulatory element", "Failure - publication describes regulatory element but there is insufficient information to annotate it", and "Failure - paper describes regulatory element but has been closed without annotation". These allow the overall success rate and failure causes to be tracked. The queue can be queried on a number of fields including user, PubMed ID, title, abstract, author, publication date, and journal. Search results can be optionally filtered by state (pending, open, or closed), TF, entry type (expert or text mining), or text-mining score. Each queue record contains a history of all state changes and comments as well as links to the publication's PubMed abstract. The current set of 'expert entry' papers in the queue was obtained from existing sources of curated publications including the Drosophila DNase I Footprint Database [8], REDfly [9], a catalog of regulatory elements for muscle-specific regulation of transcription [20, 21], ABS [4], TRED [7], ooTFD [22], DBTGR [10], or added manually by individual ORegAnno users from literature searches and review articles. The expert

entry queue currently contains 4,438 gene regulation papers of which 3,478 are open or pending and 960 are closed.

5.2.5. Development of text-mining strategies and the 'text-mining queue'

The publication queue represents an unprecedented resource for researchers interested in developing text-mining approaches to identify papers involved in gene regulation and/or extract regulatory data from these papers. In a separate study, we used both the 'success' and 'failure' papers from the 'expert-entry' to validate and compare different vector space models [23] for *cis*-regulatory document retrieval [24]. The model with the best performance in terms of sensitivity and specificity was chosen to rank the entire corpus of PubMed abstracts. By manually curating uniformly distributed samples from the top 100,000 scoring abstracts, a cut-off was set at ~58,000 so that the positive predictive value (PPV) of top scoring abstracts reached 50%, a success rate similar to that achieved during the RegCreative Jamboree (54%), and judged satisfactory by the Jamboree participants. These 58,000 papers, containing an estimated 29,000 papers that will result in regulatory annotations, have been added to the ORegAnno queue (54,351 new additions after removing duplications). We estimate that this large *cis*-regulatory text corpus will require around 15-30 person-years to be fully curated. Therefore, the Open Regulatory Annotation Consortium is actively pursuing research in text-mining techniques to identify the actual *cis*-regulatory sequences, the species, and the target gene automatically from the full text papers. In a pilot study, sequences were extracted from full-text articles for papers in the ORegAnno expert-based queue and the top 4,501 papers from the text-mining based queue. When comparing the automatically extracted data with the collection of manual ORegAnno annotations, this study achieved a reasonably high PPV (62%) at the sequence level, showing that automatic draft annotation of *cis*-regulatory elements is indeed feasible by text-mining [24]. Such draft annotations should help accelerate the manual curation and can also serve directly as benchmark data to validate *cis*-element prediction algorithms.

5.2.6. Data Access

The website (http://www.oreganno.org) provides access to an advanced search page for the entire record set, the publication queue, simple tools for scanning or extracting sequences, database dumps, and extensive help documentation. Each record page represents a complete summary of the data for a verified regulatory sequence along with links to external sources such as UCSC, Ensembl, and PubMed. All data are freely available in a number of formats without any user

registration. Users are required to register and login only if they wish to add records, comments, or scores. Nightly data dumps of the database are posted in xml format on the website. Human (hg18) and fly (dm3) records are available through the UCSC genome browser (http://genome.ucsc.edu/) as a standard track under the 'Expression and Regulation' tab. Mouse (mm8), worm (ce4) and rat (rn4) are available through the UCSC 'genome-test' browser (http://genome-test.cse.ucsc.edu/). The ORegAnno dataset is also in the process of being incorporated into the PAZAR database (760 records to date). Programmatic interaction with ORegAnno is available through web services using the Perl SOAP modules. Web services methods allow searching and retrieval of annotation records, genome mappings, and publication queue entries (see 'Dump' page for examples). Requests for the entire database (e.g. a MySQL dump) or other formats can be addressed to the authors. ORegAnno records are typically mapped to only the most current genome build for each species as provided by UCSC (e.g., hg18 for human). However, mapping can easily be performed for any other genome build upon request. A mailing list exists for updates and user assistance (oreganno@bcgsc.ca). The ORegAnno web application is available open-source under the Lesser GNU Public License at https://oreganno.dev.java.net/.

5.3. Current content of the ORegAnno database

At the time of writing, the ORegAnno database holds 30,145 records. This total includes 15,738 regulatory regions, 14,229 TFBSs, and 178 regulatory variants (polymorphisms and haplotypes) from 19 species (Table 5.3). 29,433 records have been mapped to one of 14 species representing a mapping success rate of ~98%. A large fraction of these sites were obtained from previous large-scale collections such as the FlyReg resource [8], a large set of muscle/liver-specific regulatory sites curated by Wasserman, Fickett and others [20, 25], rSNP_DB [26], a large set of human promoters [27], the REDfly resource [9], HBB and Erythroid modules [28, 29], the Vista Enhancer dataset [11], ChIP-chip sites for CTCF [30] and multiple yeast TFs [31, 32], and ChIP-Seq sites for STAT1 [33] and REST [34]. Apart from the 11 external datasets currently in ORegAnno, extensive manual curation of the literature has produced an additional 1,293 original sequence records. In total, 922 publications have been curated by 45 contributing users (from >300 registered users). The complete set of records contain regulatory sequences for 3,853 genes and 465 TFs, describe 41,856 experimental sources of evidence referencing 31 different cell types, and are further annotated by 49,807 user-comments. The majority of records (98.9%) had

positive experimental outcomes (i.e., the experiments demonstrated the sequence to be functional) but a small set of negative or neutral results have also been catalogued.

5.4. Recent Applications

The ORegAnno resource has proven useful for the development of both computational and experimental methods for the identification of novel TFBSs and regulatory polymorphisms. One such approach, called cisRED (http://www.cisred.org), uses multiple motif discovery methods applied to sequence sets that include up to 42 orthologous sequence regions from vertebrates [35]. The collection of known binding sites in ORegAnno has proved an invaluable resource for the parameter optimization and estimates of accuracy for this resource. In another study, the set of known regulatory SNPs (rSNPs) in ORegAnno was used to investigate and prioritize various properties that may be important for identifying novel regulatory polymorphisms [36]. The discriminatory potential of 23 properties related to gene regulation and population genetics was assessed by comparing these known rSNPs to a set of SNPs of unknown function (ufSNPs). A support vector machine classifier using these properties was able to discriminate rSNPs from ufSNPs with a sensitivity and specificity of 82% and 71% respectively [36]. Finally, ORegAnno has also served a critical role in the development of new experimental approaches such as ChIP-Seq. ChIP-Seq is similar to the well-described ChIP-chip method [37] except that DNA fragments isolated from the protein-DNA complex are identified by DNA sequencing instead of hybridization to a tiling microarray. The approach was first demonstrated for the STAT1 TF in interferon-γ-stimulated HeLa S3 cells [33]. A set of 41 experimentally verified sites representing 34 genomic loci for STAT1 binding were first collected from the literature and entered into ORegAnno (Oreganno dataset: OREGDS00006). Stimulated ChIP-Seq peaks were found to overlap 24 of 34 of these loci suggesting a sensitivity of ~71%. For the ORegAnno STAT1 sites shown to be functional in HeLa cells specifically, sensitivity was 100%. The collection of known STAT1 sites and binding matrices derived from them also allowed a set of high-confidence novel STAT1 binding sites to be determined and entered into ORegAnno as their own dataset (OREGDS00007). This iterative process whereby existing data drives the creation of new data demonstrates the utility and flexibility of the ORegAnno system.

5.4.1. RegCreative Jamboree

Because of its open access nature, ORegAnno has stimulated a community-wide effort for regulatory annotation. One such example, the RegCreative Jamboree

(http://www.dmbr.ugent.be/bioit/contents/regcreative/), brought together the fields of gene regulation biology, bioinformatics and biomedical text mining. There were 44 participants from 9 countries and 23 institutions. These researchers were trained to use the ORegAnno system, significantly increasing its experienced-user community. In total, 130 scientific articles were examined in depth with 96 papers meeting the criteria for annotation and resulting in 501 new regulatory sequence records. In addition to annotation, the aims of the meeting were to achieve better community standards and infrastructure for curating and storing transcriptional regulatory data, and begin to develop text mining applications for regulatory bioinformatics. It was decided that the ORegAnno evidence ontology should be made available for broader community development (see above). Similarly, the open access development and integration of TF naming conventions and sequence, cell type, cell line, and tissue ontologies were identified as future goals. The meeting also included a pre-jamboree inter-annotator agreement exercise, in which a double-blind trial was initiated prior to the jamboree to assess consistency in curation practices. This helped to identify active areas for improvement in the ORegAnno interface as well as common problems in the interpretation or reporting of regulatory sequences. We believe that the open-access data model and collaborative efforts such as the RegCreative jamboree together represent a powerful approach for regulatory annotation and continued development of the tools needed for this effort.

5.5. Conclusions

The ORegAnno resource represents the first open-access, community-based forum for annotation of regulatory sequences. ORegAnno is currently the largest collection of functionally-validated regulatory annotations available with unrestricted access. To our knowledge, it is the first resource to incorporate regulatory regions, binding sites and variation into a single resource. It is also the first system to incorporate a structured system for experimental evidence and allow both negative and positive results. The requirements for sufficient flanking sequence and verified gene identifiers (Ensembl or Entrez) ensure maximum compatibility with the community's various research needs, both currently and in the future. The intention of ORegAnno is not to replace any regulatory element databases. Many of the well-targeted databases have domain- or species-specific information that would be impractical to incorporate into a single resource. Instead, we hope to create a single multi-species database and curation system for some of the most essential information (target gene, binding protein, binding site sequence, etc.). Thus, we believe ORegAnno should exist in collaboration with the more specific databases as a central

warehouse of data, with the ultimate goal of incorporating all experimentally-verified regulatory annotation. We anticipate that this growing library of regulatory elements will prove an important resource for the validation of computational methods of motif detection, investigations of regulatory element evolution and an essential resource for the appraisal and validation of genome-wide regulatory predictions [35, 38].

Table 5.1. ORegAnno evidence classes

Evidence class	Description
Transcription regulator site	Describes evidence for the identity of a sequence that regulates transcription (e.g. transcription factor binding site).
Transcription regulator	Describes evidence for the identity of the protein that binds a transcription regulator sequence (e.g. transcription factor).
Transcript stability regulator site	Describes evidence for the identity of a sequence that regulates transcript stability.
Transcript stability regulator	Describes evidence for the identity of the protein that binds a transcript stability regulator site.
Translation regulator site	Describes evidence for the identity of a sequence (usually 3') that regulates protein translation.
Translation regulator	Describes evidence for the identity of the protein that binds a translation regulator site.

Table 5.2. ORegAnno evidence types and sub-types

Evidence Type	Evidence Subtype
UNKNOWN	UNKNOWN
Electrophoretic mobility shift assay (EMSA)	Direct gel shift
	Supershift
	Gel shift competition
	UNKNOWN
	Direct gel shift with Western blot detection
Reporter gene assay	Transient transfection luciferase assay
	Chloramphenicol acetyltransferase (CAT) Assay
	In-vivo GFP Expression Assay
	Dual luciferase reporter gene assay
	In-vivo LacZ Expression Assay
	UNKNOWN
	Temperature-sensitive mutation in transcription factor and GFP reporter construct
	Rabbit Beta-Globin expression assay
	Stable transfection luciferase assay
	Secreted alkaline phosphatase (SEAP) assay
	Transient transfection gene expression assay
	Transient transfection gene expression assay with ELISA detection
	Transient transfection GFP assay
	Targeted in-vivo LacZ expression assay by viral injection
	Site-specific stable transfection GFP assay by Cre-Lox recombination
	Transient transfection LacZ assay
	In vivo chloramphenicol acetyltransferase (CAT) assay
	In-vivo Luciferase expression assay
	In vivo gene expression assay
	In vitro virus replication assay
	Stable transfection LacZ assay
	Stably incorporated GFP transgene assay
Protein Binding Assay	Chromatin immunoprecipitation (ChIP)
	DNase Footprinting Assay
	Yeast 1-hybrid assay
	Methylation Interference Assay
	DNA-Protein Precipitation Assay
	Western Blot Assay
	Chromatin immunoprecipitation with tag sequencing (ChIP-TS)
	In-vivo Footprinting Assay
	DNA-Protein Precipitation Assay with ELISA Detection
	ChIP-on-chip (ChIP-chip)
	Chromatin immunoprecipitation with paired-end diTag sequencing (ChIP-PET)
RNA Expression Assay	RNase Protection Assay (RPA)
	Reverse Transcriptase Polymerase Chain Reaction (RT-PCR)
	Allele-specific Transcript Quantification (ASTQ)
	Competitive PCR (cPCR)
	RLM-RACE (RNA Ligase-mediated Rapid Amplification of cDNA ends)
	Whole-mount in situ hybridization
	Northern Blot
	In-vitro Transcription Assay with detection by primer extension
	Primer Extension Analysis with poly(A)+ RNA
Protein Expression Assay	Western Blot Assay
	Luciferase Expression Assay
	Enzyme-linked immunosorbent assay (ELISA)
	Indirect Immunofluorescence
Association Study	Resequencing
	Single-Stranded Conformational Polymorphism (SSCP)
	Restriction Fragment Length Polymorphism (RFLP) Analysis
RNA Stability Assay	RNA synthesis blocking
Sequence Conservation	Orthologous gene conservation
	Co-expressed gene conservation
Orthologous gene conservation	Conservation found by alignment
	Conservation found by scanning with a motif model

Evidence Type	Evidence Subtype
Linkage Analysis	Resequencing
Gene co-expression	Co-expressed genes determined through expression patterns
	Co-expressed genes determined through reporter gene experiments
	Co-expressed genes determined through microarray experiments
Mutagenesis	UNKNOWN
	Rescue construct
	Exonuclease digestion
	PCR
	Restriction endonuclease digestion
	Site-directed
	Oligonucleotide synthesis
Expert curated	Literature derived

Table 5.3. Current content of ORegAnno database

	Regulatory Haplotype	Regulatory Polymorphism	Regulatory Region	Transcription Factor Binding Site	Totals
Bos taurus			1		**1**
Caenorhabditis briggsae				21	**21**
Caenorhabditis elegans			13	194	**207**
Ciona intestinalis			7	17	**24**
Ciona savignyi			1	1	**2**
Cricetinae				3	**3**
Danio rerio			2	2	**4**
Drosophila melanogaster			680	1,415	**2,095**
Gallus gallus			8	29	**37**
Halocynthia roretzi			6		**6**
Homo sapiens	6	171	14,948	7,834	**22,959**
HIV 1				2	**2**
Mus musculus	1		55	215	**271**
Oryctolagus cuniculus				1	**1**
Rattus norvegicus			15	99	**114**
Saccharomyces cerevisiae			1	4,392	**4,393**
Takifugu rubripes				2	**2**
Xenopus laevis			1	1	**2**
Xenopus tropicalis				1	**1**
Totals (19 species)	**7**	**171**	**15,738**	**14,229**	**30,145**

Figure 5.1. Database schema (MySQL)

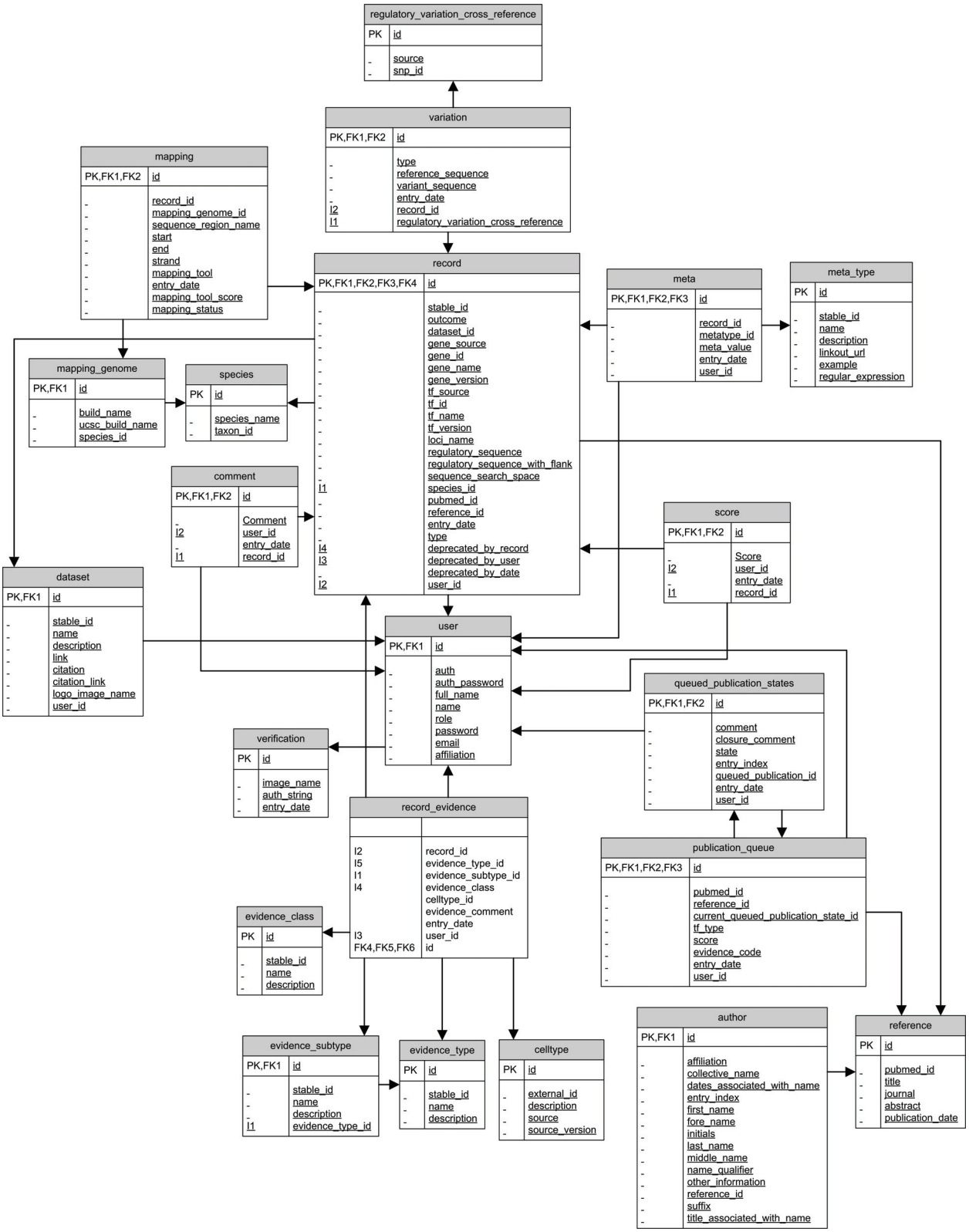
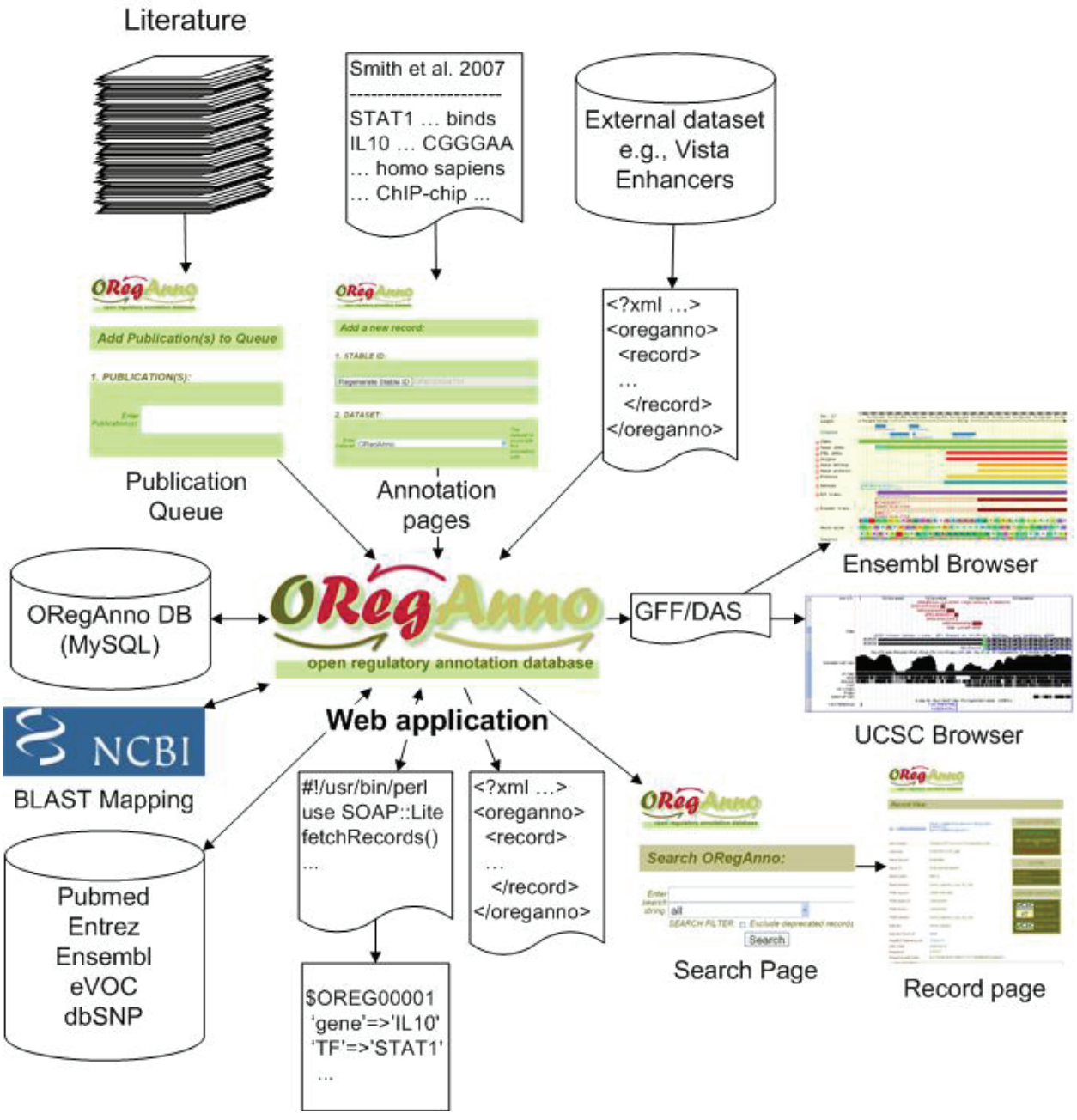


Figure 5.2. Information flow for ORegAnno annotation process

A) Data input. A publication queue allows papers from scientific literature to be added to the system for future curation. Users in the gene regulation community can enter or 'check out' papers from the queue for detailed manual curation using a series of user-friendly annotation pages. It is also possible to 'batch upload' complete datasets (e.g., external databases) using the ORegAnno XML data exchange format. B) Data storage and processing. All functionality of the ORegAnno web application depends on storage and retrieval of data from an underlying MySQL relational database. Records are cross-referenced to PubMed, Entrez, Ensembl, dbSNP, and eVOC where appropriate. A BLAST-based mapping agent assigns genome coordinates to each sequence. C) Visualization. All mapped ORegAnno records can be viewed as custom tracks in the Ensembl or UCSC genome browsers. Most records are also available as official tracks in UCSC. D) Data access. The web application provides an advanced search page for the entire record set. Each record page represents a complete summary of the data for a verified regulatory sequence. Nightly data dumps are posted in xml format. Programmatic interaction with ORegAnno is available through web services using the Perl SOAP modules.



References

1. Wasserman, W.W. and A. Sandelin, *Applied bioinformatics for the identification of regulatory elements.* Nat Rev Genet, 2004. **5**(4): p. 276-87.
2. Elnitski, L., V.X. Jin, P.J. Farnham, and S.J. Jones, *Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques.* Genome Res, 2006. **16**(12): p. 1455-64.
3. Tompa, M., N. Li, T.L. Bailey, G.M. Church, B. De Moor, E. Eskin, A.V. Favorov, M.C. Frith, Y. Fu, W.J. Kent, V.J. Makeev, A.A. Mironov, W.S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu, *Assessing computational tools for the discovery of transcription factor binding sites.* Nat Biotechnol, 2005. **23**(1): p. 137-44.
4. Blanco, E., D. Farre, M.M. Alba, X. Messeguer, and R. Guigo, *ABS: a database of Annotated regulatory Binding Sites from orthologous promoters.* Nucleic Acids Res, 2006. **34**(Database issue): p. D63-7.
5. Vlieghe, D., A. Sandelin, P.J. De Bleser, K. Vleminckx, W.W. Wasserman, F. van Roy, and B. Lenhard, *A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.* Nucleic Acids Res, 2006. **34**(Database issue): p. D95-7.
6. Matys, V., O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, and E. Wingender, *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* Nucleic Acids Res, 2006. **34**(Database issue): p. D108-10.
7. Jiang, C., Z. Xuan, F. Zhao, and M.Q. Zhang, *TRED: a transcriptional regulatory element database, new entries and other development.* Nucleic Acids Res, 2007. **35**(Database issue): p. D137-40.
8. Bergman, C.M., J.W. Carlson, and S.E. Celniker, *Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster.* Bioinformatics, 2005. **21**(8): p. 1747-9.
9. Gallo, S.M., L. Li, Z. Hu, and M.S. Halfon, *REDfly: a Regulatory Element Database for Drosophila.* Bioinformatics, 2006. **22**(3): p. 381-3.
10. Sierro, N., T. Kusakabe, K.J. Park, R. Yamashita, K. Kinoshita, and K. Nakai, *DBTGR: a database of tunicate promoters and their regulatory elements.* Nucleic Acids Res, 2006. **34**(Database issue): p. D552-5.
11. Visel, A., S. Minovitsky, I. Dubchak, and L.A. Pennacchio, *VISTA Enhancer Browser--a database of tissue-specific human enhancers.* Nucleic Acids Res, 2007. **35**(Database issue): p. D88-92.
12. Montgomery, S.B., O.L. Griffith, M.C. Sleumer, C.M. Bergman, M. Bilenky, E.D. Pleasance, Y. Prychyna, X. Zhang, and S.J. Jones, *ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation.* Bioinformatics, 2006. **22**(5): p. 637-40.
13. Hubbard, T., D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X.M. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, and E. Birney, *Ensembl 2005.* Nucleic Acids Res, 2005. **33**(Database issue): p. D447-53.

14. Maglott, D., J. Ostell, K.D. Pruitt, and T. Tatusova, *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res, 2005. **33**(Database issue): p. D54-8.
15. Wheeler, D.L., T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, J.U. Pontius, K.D. Pruitt, G.D. Schuler, L.M. Schriml, E. Sequeira, S.T. Sherry, K. Sirotkin, G. Starchenko, T.O. Suzek, R. Tatusov, T.A. Tatusova, L. Wagner, and E. Yaschenko, *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2005. **33**(Database issue): p. D39-45.
16. Sherry, S.T., M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, and K. Sirotkin, *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
17. Frith, M.C., A.S. Halees, U. Hansen, and Z. Weng, *Site2genome: locating short DNA sequences in whole genomes.* Bioinformatics, 2004. **20**(9): p. 1468-9.
18. Kelso, J., J. Visagie, G. Theiler, A. Christoffels, S. Bardien, D. Smedley, D. Otgaar, G. Greyling, C.V. Jongeneel, M.I. McCarthy, T. Hide, and W. Hide, *eVOC: a controlled vocabulary for unifying gene expression data.* Genome Res, 2003. **13**(6A): p. 1222-30.
19. Portales-Casamar, E., S. Kirov, J. Lim, S. Lithwick, M.I. Swanson, A. Ticoll, J. Snoddy, and W.W. Wasserman, *PAZAR: a Framework for Collection and Dissemination of Cis-regulatory Sequence Annotation.* Genome Biol, 2007. **8**(10): p. R207**.**
20. Wasserman, W.W. and J.W. Fickett, *Identification of regulatory regions which confer muscle-specific gene expression.* J Mol Biol, 1998. **278**(1): p. 167-81.
21. Ho Sui, S.J., J.R. Mortimer, D.J. Arenillas, J. Brumm, C.J. Walsh, B.P. Kennedy, and W.W. Wasserman, *oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes.* Nucleic Acids Res, 2005. **33**(10): p. 3154-64.
22. Ghosh, D., *Object-oriented transcription factors database (ooTFD).* Nucleic Acids Res, 2000. **28**(1): p. 308-10.
23. Glenisson, P., B. Coessens, S. Van Vooren, J. Mathys, Y. Moreau, and B. De Moor, *TXTGate: profiling gene groups with text-based information.* Genome Biol, 2004. **5**(6): p. R43.
24. Aerts, S., M. Haeussler, O.L. Griffith, S. Van Vooren, S.J.M. Jones, S.B. Montgomery, C.M. Bergman, and T.O.R.A. Consortium, *Text-mining assisted regulatory annotation.* Genome Biol, 2008. **9**(2): p. R31.1-13**.**
25. Ho Sui, S.J., J.R. Mortimer, D.J. Arenillas, J. Brumm, C.J. Walsh, B.P. Kennedy, and W.W. Wasserman, *oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes.* Nucleic Acids Res, 2005. **33**(10): p. 3154-64.
26. Ponomarenko, J.V., T.I. Merkulova, G.V. Vasiliev, Z.B. Levashova, G.V. Orlova, S.V. Lavryushev, O.N. Fokin, M.P. Ponomarenko, A.S. Frolov, and A. Sarai, *rSNP_Guide, a database system for analysis of transcription factor binding to target sequences: application to SNPs and site-directed mutations.* Nucleic Acids Res, 2001. **29**(1): p. 312-6.
27. Trinklein, N.D., S.J. Aldred, A.J. Saldanha, and R.M. Myers, *Identification and functional analysis of human transcriptional promoters.* Genome Res, 2003. **13**(2): p. 308-12.
28. King, D.C., J. Taylor, L. Elnitski, F. Chiaromonte, W. Miller, and R.C. Hardison, *Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences.* Genome Res, 2005. **15**(8): p. 1051-60.

29. Wang, H., Y. Zhang, Y. Cheng, Y. Zhou, D.C. King, J. Taylor, F. Chiaromonte, J. Kasturi, H. Petrykowska, B. Gibb, C. Dorman, W. Miller, L.C. Dore, J. Welch, M.J. Weiss, and R.C. Hardison, *Experimental validation of predicted mammalian erythroid cis-regulatory modules.* Genome Res, 2006. **16**(12): p. 1480-92.
30. Kim, T.H., Z.K. Abdullaev, A.D. Smith, K.A. Ching, D.I. Loukinov, R.D. Green, M.Q. Zhang, V.V. Lobanenkov, and B. Ren, *Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.* Cell, 2007. **128**(6): p. 1231-45.
31. Harbison, C.T., D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, J.B. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K. Pokholok, M. Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, and R.A. Young, *Transcriptional regulatory code of a eukaryotic genome.* Nature, 2004. **431**(7004): p. 99-104.
32. MacIsaac, K.D., T. Wang, D.B. Gordon, D.K. Gifford, G.D. Stormo, and E. Fraenkel, *An improved map of conserved regulatory sites for Saccharomyces cerevisiae.* BMC Bioinformatics, 2006. **7**: p. 113.
33. Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O.L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones, *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.
34. Johnson, D.S., A. Mortazavi, R.M. Myers, and B. Wold, *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.
35. Robertson, G., M. Bilenky, K. Lin, A. He, W. Yuen, M. Dagpinar, R. Varhol, K. Teague, O.L. Griffith, X. Zhang, Y. Pan, M. Hassel, M.C. Sleumer, W. Pan, E.D. Pleasance, M. Chuang, H. Hao, Y.Y. Li, N. Robertson, C. Fjell, B. Li, S.B. Montgomery, T. Astakhova, J. Zhou, J. Sander, A.S. Siddiqui, and S.J. Jones, *cisRED: a database system for genome-scale computational discovery of regulatory elements.* Nucleic Acids Res, 2006. **34**(Database issue): p. D68-73.
36. Montgomery, S.B., O.L. Griffith, J.M. Schuetz, A. Brooks-Wilson, and S.J. Jones, *A survey of genomic properties for the detection of regulatory polymorphisms.* PLoS Comput Biol, 2007. **3**(6): p. e106.
37. Ren, B., F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T.L. Volkert, C.J. Wilson, S.P. Bell, and R.A. Young, *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.
38. Xie, X., J. Lu, E.J. Kulbokas, T.R. Golub, V. Mootha, K. Lindblad-Toh, E.S. Lander, and M. Kellis, *Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals.* Nature, 2005. **434**(7031): p. 338-45.

6. Conclusions

6.1. Summary

This thesis describes the development of methods and tools for analyzing large gene expression datasets and a web resource for capturing the critical details of gene regulation experiments from scientific literature. Additionally, I describe a tissue microarray study which identified several potentially important diagnostic and prognostic biomarkers for thyroid carcinoma. Methods for combining expression and coexpression data were described and shown to increase confidence in predictions compared to single-platform or single-dataset results. Agreement between different expression datasets is often reported as disappointingly low. However, this work gives cause for optimism in that where studies do show overlap, we are left with high quality findings stripped of some of the noise that normally plagues such results. In this final chapter I will discuss some of the general conclusions that can be made from the work and suggest areas for future research.

6.2. Large-scale coexpression analysis by global and subspace methods

In our cross-platform global coexpression analysis (Chapter 2), we confirmed that intra-platform reproducibility improves with greater sample numbers and that gene pairs with higher coexpression values (e.g., Pearson correlation) are more likely to be functionally related. We also showed that combining coexpression measurements from different gene expression platforms increases confidence in those measurements as assessed by the Gene Ontology. However, in general, we observed that the level of correlation between platforms was very low, and the numbers of gene pairs that reached reasonable thresholds for ‘high-confidence coexpression’ were relatively small. As a result, I would not necessarily recommend a global coexpression approach for application to large and disparate datasets. Even greater caution is needed when combining coexpression data from different platforms based on different tissue sources and conditions. Overall, it seems that global gene coexpression (coexpression across all tissues and conditions) is relatively rare. Therefore, it may be better to approach the problem for tissue and condition subsets individually. Subspace clustering (biclustering) represents an attractive solution to this problem because it attempts to group both genes and their samples/conditions simultaneously. In this way, groups of genes are identified as coexpressed for only the tissues and samples in which coexpression is strongest. However, there are serious challenges still to overcome in the field of subspace clustering. In particular, there is a major computational barrier resulting from the nearly infinite number of possible subspaces that exist in even a modestly

sized dataset. In this work, I describe an implementation of the first subspace clustering algorithm capable of handling large gene expression datasets (Chapter 3). Biological assessment against several metrics indicates that this algorithm performs well.

Subspace clustering will become even more powerful when accurate and complete experimental annotations are available. Until recently, this has not been the case. Indeed, it is not unheard of to download a microarray dataset from a public database with no details about the experiments performed whatsoever. The researcher might have to go back to the original publications and attempt to link what was written in the methods to the samples by cryptic file names. Even when described, poor consistency in the level of detail, types of descriptions and use of non-standard terminology complicates matters. With the widespread adoption of standards such as MIAME [1], this situation has improved. However, standard ontologies for describing experiments, tissue types, and other details still need development and adoption before some of the benefits of subspace clustering can be realized. There is also room for improvement in the algorithm itself. While it is capable of running and producing results for almost any dataset, a complete solution is never achieved. The longer the program runs (determined by parameter settings), the better (more complete) the solution that is produced. Future versions should migrate the code from a graphical user interface to a command-line tool that could be parallelized. The algorithm could also be improved by allowing "ties" when building the OPSM clusters. Currently, two genes belong to the same cluster if they share the same linear ordering of values across some set of conditions. However, when genes share extremely close values for a particular experiment, the order of these values can be somewhat arbitrary. To resolve this issue, Liu and Wang (2003) suggest allowing ties within some defined threshold [2]. Basically, this allows genes to belong to the same cluster as long as they have virtually the same linear ordering of experiments. On the assessment side, it would be useful to investigate twig clusters further. These are clusters of only two genes which are tightly coexpressed across many experiments. KiWi is the first subspace clustering algorithm that can identify such clusters. However, we did not specifically evaluate the quality of these small clusters. New methods may need to be developed first as many evaluation methods (e.g., GO term over-representation) make use of statistics that do not work well for n=2.

6.3. Biomarker discovery by expression analysis, meta-analysis, and tissue microarrays

In our study of thyroid cancer we were initially presented with a large number of gene expression profiling studies from the literature reporting potential cancer biomarkers. A logical approach to this situation was to perform a meta-analysis to identify the most consistently altered genes. Ideally, this would have proceeded from raw microarray images through a consistent analysis pipeline, and produced some kind of average or cumulative statistic. However, in the vast majority of cases raw data were unavailable. Therefore, we were forced to develop a simple vote counting scheme to identify consistently reported genes from published lists of genes. Statistical analysis showed that while not ideal, this method does identify significantly over-represented genes in the data.

The 'raw data problem' is not unique to thyroid cancer and this method has proven useful for other cancers such as colon cancer [3]. The method used in this thesis was slightly improved in the colon cancer paper to give more equal weight to the fold-change and study size criteria in the ranking scheme. Subsequent analysis also showed that we perhaps unfairly over-penalize multi-study markers which had disagreements in direction (see Chapter 4). This is another area for future improvement. If this kind of meta-analysis is frequently useful, it would be advantageous to have a dynamic web resource that would allow easy updates and analyses of different comparison subsets. In such a system, users could upload lists of differentially expressed genes along with a description of the conditions compared. Then, different combinations of lists could be chosen and assessed for significant overlap. Even better would be an 'open-access oncomine' where users upload raw microarray data, facilitating the generation of true meta-analyses from raw data. This will be discussed further below in the context of my call for an 'Open Oncomine' resource.

Once the most promising markers were identified by meta-analysis, we proceeded to a tissue microarray (TMA) analysis of several of the candidates (along with a number of non-meta-analysis markers). Convincingly, all markers identified by one or more studies in the meta-analysis were significantly associated with cancer status (benign versus malignant) on the TMA and some of the higher ranking candidates were the most discriminating markers. Two thyroid cancer problems were considered: a diagnostic problem of benign versus malignant lesions; and a prognostic problem of differentiated cancer (better outcome) versus undifferentiated anaplastic cancer (very poor outcome). In both cases, we were able to identify numerous antibodies with

significantly different staining and develop classifiers with high sensitivity, specificity, and accuracy (see Chapter 4).

A number of additional analyses of the TMAs are still underway. We have ordered a number of additional meta-analysis markers that are currently being optimized and processed in the laboratory. Based on our preliminary results from only a few of the high ranking candidates, we hypothesize that these antibodies will significantly improve our classifier performance. Another TMA is also being constructed with an additional cohort of malignant and benign thyroid lesions. These will provide an important independent test set for our classifier. If the classifier passes this test, the next step will be to migrate it from archived surgical specimens to fine needle biopsy or blood samples. This will be important to show that the discriminating proteins can actually be used to diagnose malignant from benign lesions before surgery. Ultimately, this could help reduce the number of unnecessary surgeries and the associated morbidity. Similarly, the markers for (undifferentiated) anaplastic carcinoma could help identify patients at risk of progression to this more aggressive and deadly form of thyroid cancer. In addition to thyroid, our combined meta-analysis and tissue microarray analysis is being applied to several other cancers including colon, rectal, lung and breast.

6.4. Open resources for understanding gene expression and regulation

In the final main chapter (Chapter 5) of this thesis I presented ORegAnno (www.oreganno.org), a community resource for the curation of experimentally proven regulatory control sequences. This resource has proven a great success with over 30,000 sequences entered for 19 species by almost 50 contributing users. The collection has directly contributed to several publications [4-8] and is still actively growing. Near-term future developments include further integration with PAZAR [9], development of the evidence ontology, and integration of sequence, tissue and anatomical structure ontologies. We are also working on further automation of the curation process by text-mining strategies such as those developed by Aerts *et al.* (2007) [4]. ORegAnno has also recently shifted its focus slightly to the new genome-wide binding site location methods such as ChIP-chip and ChIP-seq. The 9 datasets of this type currently in ORegAnno for human and mouse probably represent the largest such collection anywhere. By allowing both low-throughput and high-throughput data types, ORegAnno facilitates a powerful approach where binding models are defined from a small set of high-quality experiments and then used to identify high-confidence binding sites in genome-wide high-throughput data. Binding sites for

several proteins of interest are currently in the pipeline using this strategy, including FoxA2 (Wederell *et al*., 2007, manuscript in preparation), E2F4, MYC, Ctcf (mouse), and HOXA9/MEIS1 as well several different histone modification types. These datasets and more should appear in ORegAnno in the near future.

Based on the success of the community-driven open-access model of ORegAnno I would like to make a call for an 'Open Oncomine' resource based on the same philosophy of community-driven data entry and free access to all data. Much like the existing Oncomine [10], this resource would collect gene expression profiling studies and their raw data and facilitate differential expression, coexpression, and meta-analyses using state-of-the-art and standardized methods. However, instead of users being able to access only a fraction of the useful data, the system would be completely opened up to them. Most cancer expression data is produced with public funds. The greatest return on the public's investment will only be realized when researchers share this data openly. The Oncomine group have done a great job of collecting large sets of expression data, creating a graphical user interface, and developing standard protocols for analysis, meta-analysis and visualization. However, they are missing the final critical components that would make this the most powerful cancer informatics tool in the world. If they wish to keep their tool closed for profit and personal gain, then we as a cancer research community must build our own.

6.5. Technological developments for gene expression and regulation analysis

In addition to the transcript-level expression profiling that was the focus of this thesis, a number of other levels of gene expression must be explored to identify relevant diagnostic or prognostic markers for cancer. In general, such investigations lag far behind transcript level profiling. Using thyroid cancer as an example, to date, only a single study has considered genome-wide differential microRNA expression [11], and none has yet applied exon arrays or alternative splicing arrays to identify cancer-specific splicing events or splice variants. Recent developments to expression profiling and sequencing technologies should help to remedy this situation. In particular, next-generation sequencing technologies hold the promise of affordable, highly sensitive, accurate, and absolute measurements of the complete transcriptome including information regarding microRNAs, alternative spliceforms, and the presence of mutations and polymorphisms. In particular, several recent studies have demonstrated the potential of using custom high-density oligo arrays to enrich DNA from targeted subsets of the genome (e.g. all

human exons) for ultra-high-throughput sequencing by next-generation sequencing technologies (e.g., by Illumina/Solexa) [12-14]. This same strategy could be used to perform sequencing of regulatory regions (e.g., all promoter regions). Microarray expression analysis of the same samples could for the first time attempt to link gene expression changes to gene regulation changes on a genome-wide scale. This together with ChIP-sequencing strategies identifying all binding sites for each transcription factor would go a long way towards a unified understanding of gene regulation. Given that gene expression changes are a fundamental feature of human cancer, such an understanding should help identify new cancer mechanisms, potential treatment targets, and have significant diagnostic and prognostic implications.

References

1. Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* Nat Genet, 2001. **29**(4): p. 365-71.
2. Liu, J. and W. Wang, *Op-cluster: Clustering by tendency in high dimensional space*, in *Proceedings of the 3rd IEEE International Conference on Data Mining*. 2003, IEEE Computer Society: Melbourne, FL, USA. p. 187-194.
3. Chan, S.K., O.L. Griffith, I.T. Tai, and S.J.M. Jones, *Meta-analysis of Colorectal Cancer Gene Expression Profiling Studies Identifies Consistently Reported Candidate Biomarkers.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(3): p. 543–52.
4. Aerts, S., M. Haeussler, O.L. Griffith, S. Van Vooren, S.J.M. Jones, S.B. Montgomery, C.M. Bergman, and T.O.R.A. Consortium, *Text-mining assisted regulatory annotation.* Genome Biol, 2008. **9**(2): p. R31.1-13**.**
5. Montgomery, S.B., O.L. Griffith, J.M. Schuetz, A. Brooks-Wilson, and S.J. Jones, *A survey of genomic properties for the detection of regulatory polymorphisms.* PLoS Comput Biol, 2007. **3**(6): p. e106.
6. Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O.L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones, *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.
7. Griffith, O.L., S.B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M.C. Sleumer, M. Bilenky, M. Haeussler, M. Griffith, S.M. Gallo, B. Giardine, B. Hooghe, P. Van Loo, E. Blanco, A. Ticoll, S. Lithwick, E. Portales-Casamar, I.J. Donaldson, G. Robertson, C. Wadelius, P. De Bleser, D. Vlieghe, M.S. Halfon, W. Wasserman, R. Hardison, C.M. Bergman, and S.J. Jones, *ORegAnno: an open-access community-driven resource for regulatory annotation.* Nucleic Acids Res, 2008. **36**(Database issue): p. D107-13
8. Montgomery, S.B., O.L. Griffith, M.C. Sleumer, C.M. Bergman, M. Bilenky, E.D. Pleasance, Y. Prychyna, X. Zhang, and S.J. Jones, *ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation.* Bioinformatics, 2006. **22**(5): p. 637-40.
9. Portales-Casamar, E., S. Kirov, J. Lim, S. Lithwick, M.I. Swanson, A. Ticoll, J. Snoddy, and W.W. Wasserman, *PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation.* Genome Biol, 2007. **8**(10): p. R207.
10. Rhodes, D.R., S. Kalyana-Sundaram, V. Mahavisno, R. Varambally, J. Yu, B.B. Briggs, T.R. Barrette, M.J. Anstet, C. Kincead-Beal, P. Kulkarni, S. Varambally, D. Ghosh, and A.M. Chinnaiyan, *Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles.* Neoplasia, 2007. **9**(2): p. 166-80.
11. Visone, R., P. Pallante, A. Vecchione, R. Cirombella, M. Ferracin, A. Ferraro, S. Volinia, S. Coluzzi, V. Leone, E. Borbone, C.G. Liu, F. Petrocca, G. Troncone, G.A. Calin, A. Scarpa, C. Colato, G. Tallini, M. Santoro, C.M. Croce, and A. Fusco, *Specific microRNAs are downregulated in human thyroid anaplastic carcinomas.* Oncogene, 2007. **26**(54): p. 7590-5.
12. Albert, T.J., M.N. Molla, D.M. Muzny, L. Nazareth, D. Wheeler, X. Song, T.A. Richmond, C.M. Middle, M.J. Rodesch, C.J. Packard, G.M. Weinstock, and R.A. Gibbs,

Direct selection of human genomic loci by microarray hybridization. Nat Methods, 2007. **4**(11): p. 903-5.
13. Okou, D.T., K.M. Steinberg, C. Middle, D.J. Cutler, T.J. Albert, and M.E. Zwick, *Microarray-based genomic selection for high-throughput resequencing.* Nat Methods, 2007. **4**(11): p. 907-9.
14. Porreca, G.J., K. Zhang, J.B. Li, B. Xie, D. Austin, S.L. Vassallo, E.M. LeProust, B.J. Peck, C.J. Emig, F. Dahl, Y. Gao, G.M. Church, and J. Shendure, *Multiplex amplification of large sets of human exons.* Nat Methods, 2007. **4**(11): p. 931-6.