Copyright

by

Patrick J. Killion

2007

The Dissertation Committee for Patrick J. Killion Certifies that this is the approved version of the following dissertation:

Fungus to Fibroblast: A Functional Genomic Exploration of Eukaryotic Transcriptional Regulation

Committee:

Vishwanath R. Iyer Supervisor

Edward Marcotte

Scott Stevens

Whitney Yin

Orly Alter

Fungus to Fibroblast: A Functional Genomic Exploration of Eukaryotic Transcriptional Regulation

by

Patrick J. Killion, B.S.

Dissertation

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

The University of Texas at Austin
December 2007

Dedication

I dedicate this work to Jenn.

She is simple in silence.

She is laughter at history.

She is most beautiful, in all ways.

To her, I will always be dedicated.

Acknowledgements

I would like to thank Vishy Iyer for providing me several opportunities in the years I have known him. My experience teaching with him at the Cold Spring Harbor DNA Microarray course catalyzed this entire process. His support from my entry into graduate research through the entire process therein has always been steady, supportive, patient, and productive. I have learned from him the value of careful thought, precise word selection, and ambassadorial treatment of colleagues.

I also thank my family for their love, understanding, patience, and strong example. My wife Jennifer is the most lovely and fun human I shall I ever know. My father has taught me the value of commitment, my word, my name, my sense of honor, and a belief in the capacity to finish. My mother has taught me of community, calm, and the value of the moments that cannot last. My sister and brother have taught me of confidence, perspective, and opportunity.

I thank the current members and alumni of the Iyer Lab. I especially thank my collaborator Dr. Zhanzhi (Mike) Hu for the several years we spent working on *The Network*.

Thank you to my committee members for their time, efforts, and concern. I would like to especially thank Dr. Scott Stevens for the balanced servings of professional advice and personal friendship he has offered for many years. I have also greatly valued the weekly insights and ideas of Dr. Edward Marcotte in our joint lab meetings.

Fungus to Fibroblast: A Functional Genomic Exploration of Eukaryotic Transcriptional Regulation

Publication No._____________

Patrick J. Killion, Ph.D.
The University of Texas at Austin, 2007

Supervisor: Vishwanath R. Iyer

I have pursued a breadth of research that explored the functional genomic study of eukaryotic transcriptional regulation. I have utilized two model organisms, many experimental methodologies, and have developed a suite of computational resources to study the interaction of transcription factors with regulated targets.

In *Saccharomyces cerevisiae* I worked with my collaborator Dr. Zhanzhi (Mike) Hu to characterize the whole-genome transcriptional response of 263 individual transcription factor deletions. We utilized a sophisticated error model and directed-weighted graphs to model a network of high-confidence targets for each transcription factor profiled. We then used regulatory epistasis to elucidate the true set of primary KO-regulated targets and construct a functional transcriptional regulatory network. This network was analyzed for ontological and sequence motif enrichment in order to gain insight into the biological functions represented by transcription factors studied. Functional validation was performed to evaluate the probability of novel functional

characterizations. Significant insight was gained from this study with regard to the nature of regulatory cascades and the inability for DNA binding events to predict regulation.

This set of analysis was performed with a novel bioinformatic server called ArrayPlex. ArrayPlex is a software package that centrally provides a large number of flexible toolsets useful for functional genomics including microarray data storage, quality assessments, data visualization, gene annotation retrieval, statistical tests, genomic sequence retrieval and motif analysis. It uses a client-server architecture based on open source components, provides graphical, command-line, as well as programmatic access to all needed resources, and is extensible by virtue of a documented API.

Using many of the techniques and computational resources developed, I pursued the study of microRNA transcriptional abundance and targeting in *H. sapiens* cell cultures. Utilizing custom-fabricated microarrays, I measured the whole-genome response of both mRNAs and microRNAs under serum stimulation, c-Myc overexpression, and c-Myc siRNA-mediated knockdown. I then characterized the regulatory interactions between the sets of regulated microRNAs and coordinately regulated transcription factors. Using analytical methods sensitive to regulatory directionality of both populations I was able to determine high-confidence relationships between transcription factors and regulated microRNAs as well as microRNAs and regulated gene targets.

Table of Contents

List of Figures xiv

List of Tables xvi

Chapter 1: Introduction 1
  - Eukaryotic Transcriptional Regulation 1
    - Complexity Through Regulation 1
    - Eukaryotic Transcriptional Regulation 1
    - MicroRNA Regulation 3
  - Research Overview 5

Chapter 2: General Materials and Methods 9
  - DNA Microarray Fabrication 9
    - Yeast Probe Set 9
    - Human mRNA Probe Set 9
    - Human MicroRNA Probe Set 9
    - Printing 9
  - Primary Data Capture & Transformation 10
    - Microarray Scanning 10
    - Data Warehousing 10
    - Data Normalization 10
      - Positive Control Normalization 11
      - Global Normalization 11

Chapter 3: Bioinformatic Infrastructure 12
  - Introduction 12
  - Results 13
    - Primary Data Warehousing 13
      - Database Selection 14
      - The Open-Source Translation 16

      - MIAME Support 17
    - Final Release & Support 19
    - Secondary Data Association & Analysis 19
    - Core Technology, Design, & Network Operation 20
      - Bundled Genomic Resources 23
      - Integrated Sequence Analysis Toolsets 29
      - Analytical Accessibility & Customization 30
    - Documentation and Guidance 35
  - Discussion 36
    - Utility of the Primary Data Warehouse...36
    - LAD Adoption at Research Institutions 36
    - ArrayPlex Analytical Proving Ground...37
      - High-throughput Microarray Data Quality Analysis 37
      - Ontological Enrichment & Connectivity 41
      - Visualization Capacity 43
    - Comparison to Similar Software Packages 43
  - Materials and Methods 44
    - Longhorn Array Database Requirements & Availability 44
    - ArrayPlex Requirements & Availability 44
    - ArrayPlex Server Requirements 44
    - ArrayPlex Client Requirements..45
    - ArrayPlex Command-Line Module Requirements 46
  - Supplemental Information 46
    - ArrayPlex Client Feature Set 46
- Chapter 4: A Functional Transcriptional Regulatory Network 47
  - Introduction 47
  - Results 48
    - Primary Data Processing 48
      - Bias Screen 49

  - Strain, Growth & Microarray Normalization 49
  - Secondary Data Determination by Error Model 49
  - Deletion Expression Validation 53
  - Correlative Target Overlap 55
- The Functional Regulatory Network 55
  - Tertiary Data Model 56
  - The Regulatory Network G(N) 61
  - Regulatory Cascade Detection 61
  - Regulatory Cascade Results...62
- Ontological Enrichment Analysis...66
  - Gene-Specific Assignment Acquisition 66
  - Ontological Network Modeling 68
  - Raw Term Enrichment 71
  - Composite Term Enrichment...72
  - Enrichment Results 74
  - Term Enrichment Validation 74
- Network Analysis of Enriched Terms...78
  - Network Metric, Average Path Length 80
  - Network Metric, Clustering Coefficient...................................82
  - Network Analysis Results..84
- Target Promoter Sequence Motif Analysis 84
  - Sequence Analysis Life Cycle 85
  - Promoter Sequence Extraction...87
  - Motif Search..87
  - Motif Search Results 88
  - Motif Discovery 90
  - Motif Discovery Results & Aggregation..................................90
  - Final Motif Aggregation...91
- Binding Does Not Affirm Regulation..92
- Factor on Factor Regulation...94

Discussion 97
Materials and Methods 100
    Experimental Conditions...100
        Strain Information 100
        Non-Essential Transcription Factor Growth Conditions..........100
        Essential Transcription Factor Growth Conditions.................100
        Heat Shock Growth Conditions 100
        Overexpression of HSF1 101
        Growth Defect Assay 101
    DNA Microarray Methods 101
        Total RNA Isolation 101
        Reverse Transcription...102
        cDNA Fluorophore Labeling 102
        Microarray Slide Preparation...102
        Hybridization & Washing..103
    Computational Methods 103
        Network Modeling 103
        GEO Repository..103
        Longhorn Array Database..103
Supplemental Information 104
Chapter 5: MicroRNA Transcriptional Abundance & Regulation 106
Introduction..106
Results 107
    Experimental Design...107
    MicroRNA Microarray Design...108
        Initial Limitations..108
        Labeling Methods..111
        Slide Substrate..111
        Updated MicroRNA Probe Set 112

        Development of a Reliable Process..112
    Primary Data Normalization...115
        Positive Control Normalization 115
        Global Normalization 116
        Normalization Conclusion 116
    Regulation of MicroRNAs by Serum Stimulation...........................119
        Experiment Design...119
        Baseline Tests 121
        Novel MicroRNA Targets 125
        Candidate Regulators 128
        Candidate Regulator Motif Search...130
        Generalized Motif Search..131
        mRNA Targeting..133
        Ontology Enrichment for Target Sets 141
    Regulation of MicroRNAs by c-Myc 142
        Experiment Design...142
        Baseline Tests 144
        Novel MicroRNA Targets 147
        General E-box Search..150
        Extended E-box Analysis 152
        c-Myc MicroRNA Intermediaries...156
        Candidate Cofactors 160
    Real-Time PCR Validation 162
Discussion..164
    Process Determination 164
    Filtered Term Enrichment...165
    Sequence Alignment 167
Materials and Methods 168
    Cell Culture & Experiments..168
        Normal Cell Culture Conditions 168

        2091 Fibroblast Serum Stimulation..168
        c-Myc Overexpression Lipotransfection 169
        c-Myc siRNA Lipotransfection...169
    General DNA Microarray Methods 171
        Total RNA Isolation 171
    mRNA Expression Microarray Methods 171
        Reverse Transcription...171
        cDNA Fluorophore Labeling 171
        Microarray Slide Preparation...171
        Hybridization & Washing..172
    MicroRNA Expression Microarray Methods 172
        Synthesis of Positive Control RNA..172
        MicroRNA Enrichment, RNA Size Fractionation 173
        Fluorophore Labeling, Poly-A Method 173
        Fluorophore Labeling, ULS Method 173
        Microarray Slide Preparation...173
        Hybridization & Washing..174

Chapter 6: Conclusions and Recommendations 175
    Bioinformatic Infrastructure 175
    MicroRNA Transcriptional Abundance 177

Appendix I – ArrayPlex Client Feature Set..180

Appendix II – GO Slim Term Enrichment...191

Appendix III – Novel Sequence Motifs 199

References 210

Vita 217

List of Figures

Figure 1.1 – MicroRNA Biogenesis.......8
Figure 3.1 – Longhorn Array Database Core Technology.......18
Figure 3.2 – ArrayPlex Core Technology.......22
Figure 3.3 – ArrayPlex Architecture, Resources, & Communication.......24
Figure 3.4 – ArrayPlex Client & Server API Network Operation.......33
Figure 3.5 – ArrayPlex Client & Module Pairing.......34
Figure 3.6 – GenePix Result File Batch Quality Screen.......39
Figure 3.7 – Ontological Assignment Enrichment & Connectivity.......42
Figure 4.1 – Error Model vs. Arbitrary Filter Cutoffs.......52
Figure 4.2 – Deletion Validation.......54
Figure 4.3 – Sample Interaction File.......58
Figure 4.4 – Network Modeling by *P Value*.......59
Figure 4.5 – Network Modeling by Weighted X Score.......60
Figure 4.6 – Network Refinement Process.......63
Figure 4.7 – Network Refinement Example.......64
Figure 4.8 – The Regulatory Network *G*(*RN*).......65
Figure 4.9 – Ontology & Genome Sequence Retrieval.......67
Figure 4.10 – GO Ontology Network Modeling.......69
Figure 4.11 – Transcription Factor Ontology Network.......70
Figure 4.12 – Composite GO Analysis.......73
Figure 4.13 – Condition-Specific GO Term Enrichment Validation.......77
Figure 4.14 – Average Path Length.......81

Figure 4.15 – Clustering Coefficient with GO Term Enrichment 83
Figure 4.16 – Sequence Analysis Life Cycle 86
Figure 4.17 – Sample Motif Search & Discovery Results 91
Figure 4.18 – Factor on Factor Visualization 95
Figure 5.1 – Initial MicroRNA Microarray Signal Bias 110
Figure 5.2 – ULS-labeled MicroRNA Microarray 113
Figure 5.3 – MicroRNA Microarray Procedural Flow 114
Figure 5.4 – Normalization Method Comparison 118
Figure 5.5 – Serum Stimulation Experiment Design 120
Figure 5.6 – Activated MicroRNA Comparison 122
Figure 5.7 – Activation of mir-17/mir-106 Clusters 123
Figure 5.8 – Activation of MicroRNA mir-34a 124
Figure 5.9 – Serum Stimulation MicroRNA Targets (Normal) 126
Figure 5.10 – Serum Stimulation MicroRNA Targets (Restricted) 127
Figure 5.11 – Candidate Regulator Identification 129
Figure 5.12 – Serum Stimulation mRNA Targeting 137
Figure 5.13 – c-Myc Overexpression & siRNA Experiment Design 143
Figure 5.14 – c-Myc Modulation mir-17/mir-106 Cluster Effect 145
Figure 5.15 – Cell Line mir-17/mir-106 Cluster Effect 146
Figure 5.16 – Novel c-Myc MicroRNA Targets 149
Figure 5.17 – Extended E-box Analysis 154
Figure 5.18 – cMyc MicroRNA Intermediaries 158
Figure 5.19 – Candidate Cofactors 161
Figure 5.20 – Real Time PCR Validation 163

List of Tables

Table 3.1 – ArrayPlex Managed Resources 25

Table 3.2 – ArrayPlex Integrated Toolsets 26

Table 3.3 – ArrayPlex Modules 27

Table 4.1 – GO Term Enrichment For Growth Assays 76

Table 4.2 – Graph Definitions for GO Network Analysis 79

Table 4.3 – Complete Motif Search Results 89

Table 4.4 – HSF1 Post-Binding Regulatory Analysis 93

Table 4.5 – Transcription Factors Profiled 104

Table 5.1 – Predicted & Known Regulated MicroRNA Targets 138

Table 5.2 – Predicted & Measured Regulated MicroRNA Targets 139

Table 5.3 – Predicted & Multi-Source Regulated MicroRNA Targets 140

Table 5.4 – General E-box Search Results 151

Table 5.5 – Extended E-box Analysis Results 155

Table 5.6 – Regulated Gene Targets of c-Myc Regulated MicroRNAs 159

Chapter 1: Introduction

EUKARYOTIC TRANSCRIPTIONAL REGULATION

Complexity Through Regulation

Modern comparative genomic analyses indicate that the physiological complexity exhibited by metazoans is not accounted for through what was once predicted would be a proportionate increase in total gene count [1]. *C. elegans* has a 97 Mb genome of approximately ~19,000 genes while the physiologically more complex *D. melanogaster* has a 180 Mb genome containing only ~13,600 genes [2, 3]. The 3,200 Mb *H. sapiens* genome, containing ~30,000 protein-coding genes, was surprisingly smaller than even contemporary expectations [4]. Alternative splicing, the production of multiple mRNA isoforms from a single genomic sequence, is posited as one significant source of higher-eukaryotic diversity [5]. Indeed, recent analysis of *C. elegans*, *D. melanogaster*, and *H. sapiens* genome sequences has estimated alternatively spliced isoforms at 5%, 18%, and 75% of putative genes, respectively [1, 3, 6]. In additional to spliceosomal variation, more sensitive control of gene expression and regulation has become increasingly implicated in the characterization of cellular complexity and differentiation [7, 8].

Eukaryotic Transcriptional Regulation

Within the context of a eukaryotic cell, only a small percentage of genes in a genome are being expressed at any given time. The process of regulating which genes are actively transcribed is a direct requirement of the diverse cellular types, physiological states, and environmental conditions that define an organism and its ecosystem.

Transcriptional regulation is facilitated by many interdependent and parallel processes: chromatin modification, sequence-specific DNA-protein transcription factor interactions, as well as the managed assembly of general eukaryotic transcriptional machinery all determine the expression level of a gene at any given time [7].

Core to the process of differential gene expression and cell-fate determination are the cis-regulatory elements that define eukaryotic promoters. Eukaryotic promoters generally contain three common elements, the first two defining what is often called the core promoter: the transcription-start site, the TATA box, and upstream sequences including activators, enhancers, repressors, and silencers [9]. The core promoter, extending approximately 100bp and including the transcription start site, is generally adequate for directing the initiation of a basal level of ubiquitous transcription, thereby providing function but lacking in regulatory potential[7]. A subset of eukaryotic core promoters contain a TATA box approximately 30bp upstream of the transcription start site [9, 10].

Contemporary molecular biology now encompasses both the painstaking characterization of just a few genes at a time as well as the *in vivo* whole-genome transcriptional expression profiling of eukaryotic experimental samples under a variety of cellular and environmental conditions [11-14]. The data produced by these experiments hardly make the regulatory mechanisms of cellular systems transparent, however. Nonetheless, the combination of whole-genome expression data with the modern availability of sequence annotation has accelerated progress towards the characterization of biological systems as deterministic module networks [15]. Expression profiles provide both an indication of temporal interaction amongst putative genes in a network as well as a direct means of testing complex regulatory relationships through the transcriptional characterization of deletions, mutants, or RNAi-mediated knockdowns. These techniques, however, have not been able to elucidate the entire spectrum of regulation operating within eukaryotic organisms. There are certainly known mechanisms of control that cannot be accounted for by expression profiling, including regulated post-transcriptional activation [16]. Perhaps, however, clear understanding has been lacking due to an incomplete tally of the players and their roles within the game.

MicroRNA Regulation

Over a decade ago, the *C. elegans* heterochronic gene *lin-4* was discovered to be required for proper post-embryonic development [17]. Three features of *lin-4*'s characterization was the discovery that it codes for a 22nt non-coding RNA (ncRNA) rather than a protein product, that its mature form contains partial antisense complementarity to the protein-coding gene *lin-14*, and that this complementarity eventually guides the post-transcriptional degradation of *lin-14* [18]. Some time passed before the characterization of *C. elegans let-7* showed that the regulatory phenomenon was not a singular aberration [19]. It has since become clear that *let-7*'s similar sequence-directed degradation of protein-coding transcripts is highly conserved from *C. elegans* to *H. sapiens* and that this widespread regulatory phenomena of microRNAs has both significant representation and conservation within nearly all metazoan genomes [18].

MicroRNA biogenesis has only recently become well characterized (Figure 1.1). Nuclear transcription and pre-processing of endogenous microRNA genes, as facilitated by *Drosha*, creates pri-microRNAs that are quickly processed into ~70nt pre-microRNAs [18, 20]. Pre-microRNAs all have primary sequences that ensure a stem-loop secondary structure [20]. Pre-microRNAs are then exported from the nucleus to the cytoplasm by Exportin (Exp5/Ran-GTP) mediated cargo transport [21]. Once present in the cytoplasm, pre-microRNAs are again processed by the well-conserved protein complex Dicer. The involvement of Dicer in microRNA processing is similar to its characterized role in the processing of double-stranded RNA during RNA interference (RNAi) [22]. Mature microRNAs now interact with and direct multi-protein RISC complexes, which mediate either the translational repression or post-transcriptional cleavage of protein-coding transcripts [20]. In this manner microRNAs have a substantial post-transcriptional effect upon the overall regulation of gene expression.

If microRNA-directed regulation is an essential component of the regulatory fabric that defines metazoan development and homeostasis, the expression of microRNAs themselves must indeed be carefully modulated. Additionally, if the original transcription of microRNAs is mediated by RNA polymerase II, well-tested computational and experimental methods will provide clues to mechanisms that define their regulation.

Though a few microRNAs have been mapped within intronic regions of known genes, microRNAs generally reside within the great expanses of intergenic space that are present in higher eukaryotes [23]. Very few of the promoter regions of intergenic microRNAs have been characterized. Nonetheless, significant evidence supports the hypothesis that these microRNAs are the product of RNA polymerase II thereby implicating regulation by cis-acting sequences and trans-acting factor [24, 25]. Specifically, a documented cluster of *C. elegans* microRNAs (*mir-35-mir-41*) show transient patterns of expression during embryogenesis specifically suggesting regulated transcription [26]. Recent phylogenetic and molecular study of microRNAs in *C. elegans* and *D. melanogaster* have identified transcriptional responses to extracellular signaling, conserved promoter elements, and trans-acting factors [27]. Additionally, several research groups have both located and directly implicated the combination of specific

DNA sequences and transcription factors with modulated mammalian microRNA expression [28-39]. Given the fundamentally powerful yet almost entirely uncharacterized role microRNAs play within the context of whole-genome regulation and environmental response, it is important that the mechanisms, sequences, and factors directing their own regulated expression be properly characterized. Finally, as microRNAs become increasingly implicated in causal roles of a large variety of human disease conditions, it is critical that our understanding of regulatory interactions that mediate their transcriptional regulation become more understood.

RESEARCH OVERVIEW

I have pursued research that explored the functional genomic study of eukaryotic transcriptional regulation. I have utilized two model organisms, many experimental methodologies, and have developed a suite of computational resources to study the interaction of transcription factors with regulated targets.

Chapter 2 is a succinct summation of generalized DNA microarray fabrication, primary data production, and normalization methods that are globally applicable to all research discussed.

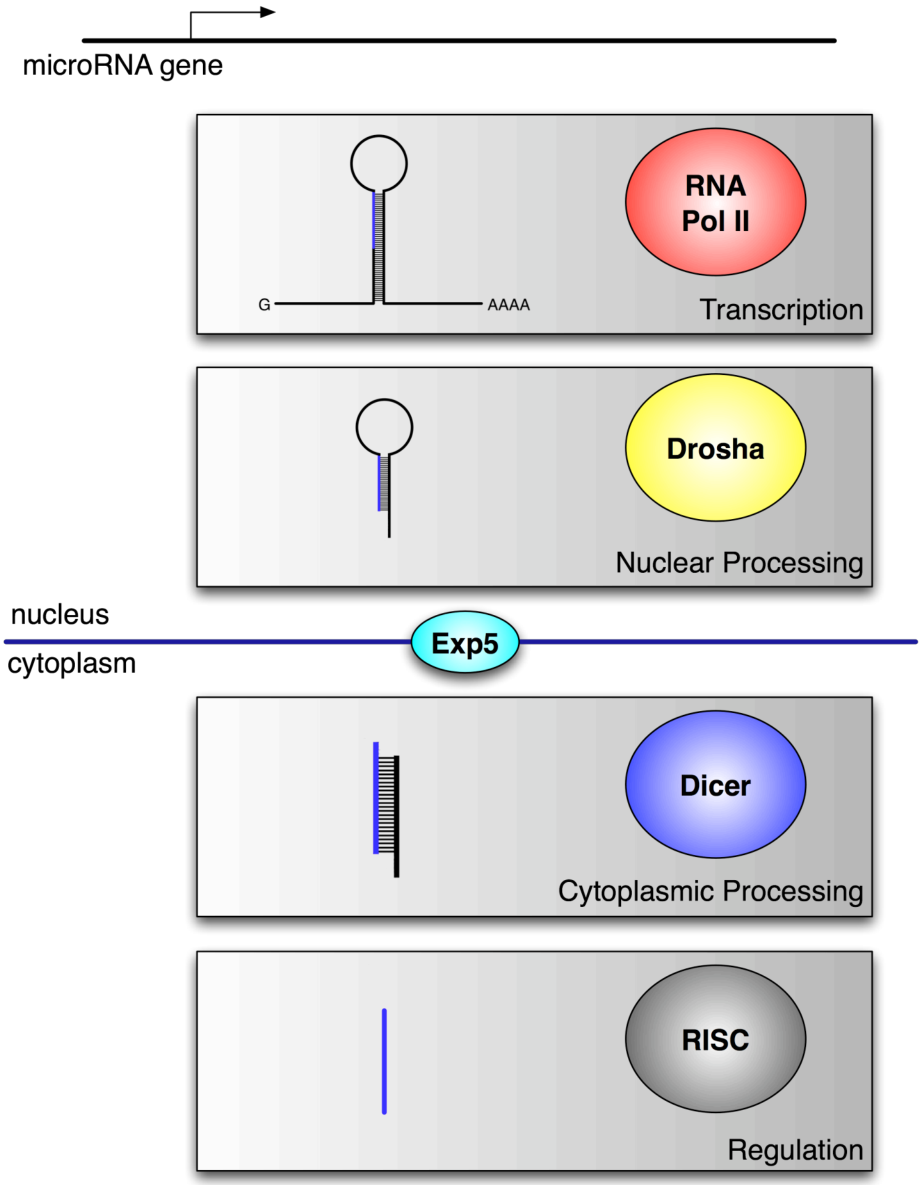
Chapter 3 details the rationale and methods behind the development of bioinformatic infrastructure resources. The Longhorn Array Database is well known at the University of Texas at Austin as the central warehouse for nearly all locally produced primary microarray data. ArrayPlex is a software package that centrally provides a large number of flexible toolsets useful for functional genomics including microarray data storage, quality assessments, data visualization, gene annotation retrieval, statistical tests, genomic sequence retrieval and motif analysis. It uses a client-server architecture based on open source components, provides graphical, command-line, as well as programmatic access to all needed resources, and is extensible by virtue of a documented API.

Chapter 4 details my work with Dr. Zhanzhi (Mike) Hu to characterize the whole-genome transcriptional response of 263 individual transcription factor deletions. We utilized a sophisticated error model and directed-weighted graphs to model a network of high-confidence targets for each transcription factor profiled. We then used regulatory epistasis to elucidate the true set of primary KO-regulated targets and construct a functional transcriptional regulatory network. This network was analyzed for ontological and sequence motif enrichment in order to gain insight into the biological functions represented by transcription factors studied. Functional validation was performed to evaluate the probability of novel functional characterizations. Significant insight was gained from this study with regard to the nature of regulatory cascades and the inability for DNA binding events to predict regulation.

Chapter 5 details my study of microRNA transcriptional abundance and targeting in *H. sapiens* cell cultures. This research was a final important step in the goals I had with respect to my progress as a biologist and study of eukaryotic transcriptional regulation. First, it was a project in which I assumed complete leadership of process development, experiment design, and data analysis. MicroRNA expression profiling had been prototyped in the Iyer Lab but was far from a standardized process with Dr. Jian Gu graduated. I was eager to accept the challenge of making the experimental process work in a deterministic and reliable manner. The enrichment, direct labeling, and hybridization of small quantities of RNA proved to be a non-trivial methodological challenge. Utilizing custom-fabricated microarrays, I measured the whole-genome response of both mRNAs and microRNAs under serum stimulation, c-Myc overexpression, and c-Myc siRNA-mediated knockdown. I then characterized the regulatory interactions between the sets of regulated microRNAs and coordinately regulated transcription factors. Using analytical methods sensitive to regulatory directionality of both populations I was able to determine high-confidence relationships between transcription factors and regulated microRNAs as well as microRNAs and regulated gene targets.

Figure 1.1 – MicroRNA Biogenesis

The process of microRNA biogenesis begins in the nucleus with microRNA gene transcription by RNA polymerase II to generate pri-microRNAs. The long initial transcripts are processed by Drosha/DGCR8 into ~70nt pre-microRNAs. Exportin5 exports the pre-microRNAs. In the cytoplasm Dicer processes the pre-microRNA into a microRNA duplex. The duplex is separated and the single-stranded mature microRNA associates with RISC complex.



Chapter 2: General Materials and Methods

DNA MICROARRAY FABRICATION

This study required the production of three types of custom DNA microarrays. All DNA microarrays were made in the Iyer Lab through the collective efforts of graduate students, post-doctoral researchers, and adviser oversight.

Yeast Probe Set

The yeast probe set consisted of cDNA sequences produced by PCR amplification of all yeast ORF and intergenic sequences.

Human mRNA Probe Set

The human mRNA probe set consisted of purchased cDNA sequences from 47,000 sequence-verified IMAGE clones.

Human MicroRNA Probe Set

The human microRNA probe set consisted of Ambion DNA oligonucleotides representing 281 *H. sapiens* mature microRNAs, 49 *M. musculus* mature microRNAs, 14 *R. norvegicus* mature microRNAs, and several dozen negative and positive control sequences.

Printing

DNA microarrays were made by using a customized robotic arrayer to spot DNA on either produced poly-L-lysine coated slides (yeast) or CEL VEPO-25C Epoxy Vantage Slides epoxy slides (Human mRNA, microRNA).

PRIMARY DATA CAPTURE & TRANSFORMATION

Microarray Scanning

Molecular Devices GenePix Pro 5.0 and 4000A/4000B microarray scanners were used to scan all DNA microarray slides. Fluorescent intensities of hybridized samples were measured on either the Cy5 (635nm) or Cy3 (532nm) responsive wavelengths. Primary data was saved in the form of uncompressed TIFF files for each fluorescent wavelength. Primary TIFF files were annotated through the process referred to as *gridding*. Circles were aligned with fluorescent spots. Each circle was pre-annotated to correspond with a documented DNA sequence. The process of *gridding* was saved as a GenePix Pro Settings file (GPS). Final data generation produced more than 50 metrics per annotated spot in a GenePix Pro Results file (GPR) that was uploaded to the Longhorn Array Database.

Data Warehousing

The Longhorn Array Database, described in Chapter 3, was used to store all primary microarray data from all microarray experiment types.

Data Normalization

The process of data normalization did not modify pre-normalized primary data files and values. Normalized values were determined for all applicable single-channel and ratio-based metrics and stored alongside unmodified primary data in the database.

Positive Control Normalization

Positive control normalization involved the manual application of a predetermined normalization coefficient to an experiment during the process of Longhorn Array Database experiment submission. This normalization coefficient was determined by custom software that read a GenePix Pro Results file and determined a linear coefficient that would adjust the average positive control spot ratio to a value of 1.0.

Global Normalization

Global mean normalization was applied to all experimental results during the process of data submission to the Longhorn Array Database with the exception of when positive control normalization was manually applied. Global mean normalization involved the computation of a linear normalization coefficient that would adjust all spot ratios such that their mean value was 1.0. This coefficient was determined by the Longhorn Array Database and applied without user intervention or calculation.

Chapter 3: Bioinformatic Infrastructure

The efforts described in this chapter detail the rationale and implementation of two bioinformatic systems: Longhorn Array Database (LAD) and ArrayPlex. The former project was in progress in 2002 and continued through 2003. The research article *The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD)* was published in August of 2003 in *BMC Bioinformatics* [40]. A book chapter entitled *Microarray Data Visualization and Analysis with the Longhorn Array Database (LAD)* was authored shortly thereafter for *Current Protocols in Bioinformatics* [41]. ArrayPlex was initiated in 2004 and was instrumental in the fulfillment of research efforts described in Chapters 4 and 5. ArrayPlex is an independent manuscript (*ArrayPlex: distributed, interactive and programmatic access to genome sequence, annotation, ontology, and analytical toolsets*).

INTRODUCTION

Microarray data analysis is divided into two distinct stages. The first stage involves the warehousing of primary data into a centralized repository. This takes the form of a database system that has been specifically designed to handle DNA microarray data. A database system is requisite to the process of even the simplest of primary data analysis procedures. The second stage of analysis is one of external information association and biological discovery. Primary microarray data that has been filtered for both quality and significance is the beginning of the process of taking raw measurements and elucidating biological meaning. That process cannot continue without the association of standardized gene annotations, ontological associations, and genome sequence with the microarray spots that emerged from the primary analysis and filtering process.

Each of these analytical stages requires specialized bioinformatic systems. My initial research efforts began with the extended customization and deployment of an existing microarray database system. This system, the Longhorn Array Database (LAD) would prove to be the foundation of both my personal research and the primary microarray database for more than one hundred Iyer Lab members and collaborators. Secondarily, I designed and implemented a novel network-centric computational environment focused entirely upon centralized access to genome resources.

RESULTS

Primary Data Warehousing

A microarray database fulfills specific needs for a research environment that produces even modest quantities of primary microarray data. Its primary function is to parse the file-format of the primary data itself and store the contained information in a high-performance underlying relational database system. The relational system is usually one of several open-source or commercial variants: MySQL, PostgreSQL, or Oracle. Once stored in the database system, a single or group of microarray experiments can be uniformly filtered for both data quality and statistical significance. The single most important function of the microarray database is to accurately execute this process of database population with each experiment submission and then to subsequently protect the primary data from corruption. Technologically, this protection comes in several forms. Relational databases themselves have a feature-set known as transactions that allow operations that modify or delete data to be bundled in an all-or-nothing execution package. If an error occurs or data inconsistency is detected during the process of data manipulation the database will rollback the entire set of operations to the previous error-free state. This feature is critical to protecting primary data but can be utilized only if the microarray database application itself makes use of the option to do so.

Microarray databases secondarily fulfill the need for geographic and collaborative flexibility in the process of data submission and access. Centralized location and network accessibility allow groups of users to both submit primary data and share access to result sets and analytical toolsets.

Database Selection

We evaluated many relational database packages developed specifically for the purpose of primary DNA microarray data storage. ArrayDB, BASE, GeneX, MADAM and MIDAS were all examples of development efforts from academic sources [42, 43]. Most of these solutions had the advantage that the system source code was available and were thus open to modification. Many of them had significant shortcomings. Several stored numerical data but lacked the ability to simultaneously archive and visualize primary array images. Others were simply not capable of operating with thousands of loaded experiments or were built upon technologies that would not guarantee the integrity of the stored primary data. Finally, some solutions did not provide for network-based remote user access, a functional necessary in a collaborative multi-researcher environment. GeneX, for example, was designed to rely upon a suite of both web-based and local applications for a variety of import and analysis feature. This architectural decision expanded its graphical feature potential yet simultaneously reduced its geographical and web distributed user capabilities important to many research environments.

One of the most utilized microarray databases was the Stanford Microarray Database [44]. During the period in which we evaluated microarray database options, SMD had already archived more than 34,500 microarray experiments including 4500 from more than a hundred different publications and supported approximately 700 users. It had a great breadth of features that included data filtering, data analysis, visualization toolsets, and updated gene annotations for many model organisms. Additionally, SMD featured a strict hierarchical user and group model of user accounts such that experiments could be collaboratively shared or protected.

The source code for SMD had been open-source for some time, which theoretically allowed any researcher to install and operate an SMD server. SMD in this form, however, was based on proprietary hardware and software infrastructure that would have required a significant capital expenditure from any laboratory that wished to operate such a server. SMD was primarily operated on the Sun Solaris operating system. Solaris was tailored for Sun hardware and processors that were incompatible with and not nearly as affordable as Intel-based computer systems. Additionally, SMD was designed and written to utilize the Oracle relational database management system. The cost of initial investment and long-term ownership of these technologies was significantly higher than alternative open-source technology choices. In addition to licensing costs, Oracle is a very demanding database in terms of the expertise required by professional database administrators to deploy and maintain it.

Given the numerous strengths and proven nature of SMD, we wished to adapt it for compatibility with open-source operating and relational database systems. We chose the combination of Linux and PostgreSQL to replace Solaris and Oracle. We have named our open-source version of SMD the Longhorn Array Database (LAD).

The Open-Source Translation

We adopted a two-step strategy to accomplish the open-source port. We first transitioned SMD from Solaris to Linux while using Oracle as the relational database. This allowed us to ensure that the application component of the system operated to expectations before introducing major changes to the application source code. We conservatively introduced only the required changes throughout the SMD/LAD source tree such that it would reliably operate within the Linux operating system.

Once SMD/LAD was compatible with Linux, we undertook its migration to an open-source relational database that could support all the features required by the database-specific command-set within the SMD/LAD application source code. These included transactions, foreign-key integrity constraints, indexes, and sequences. The open-source relational database that met these requirements was PostgreSQL. PostgreSQL is an open-source object-relational database management system that supports nearly all SQL constructs including transactions, triggers, stored procedures, sub-selects, and user-defined types and functions. The use of such features is generally considered critical for ensuring data integrity. MySQL, another open-source relational database system evaluated, had only attained much of this feature set and was not known to operate at the scale intended for our microarray database. The original SMD database schema was re-implemented in the format required by PostgreSQL. Oracle-specific SQL code, constructs, and syntax in the SMD tree were translated to a standards-compliant SQL set of statements so that they would execute correctly with PostgreSQL. We optimized the indexing of certain tables and profiled query execution of involved multi-table joins to ensure that operations would complete in acceptable timeframes.

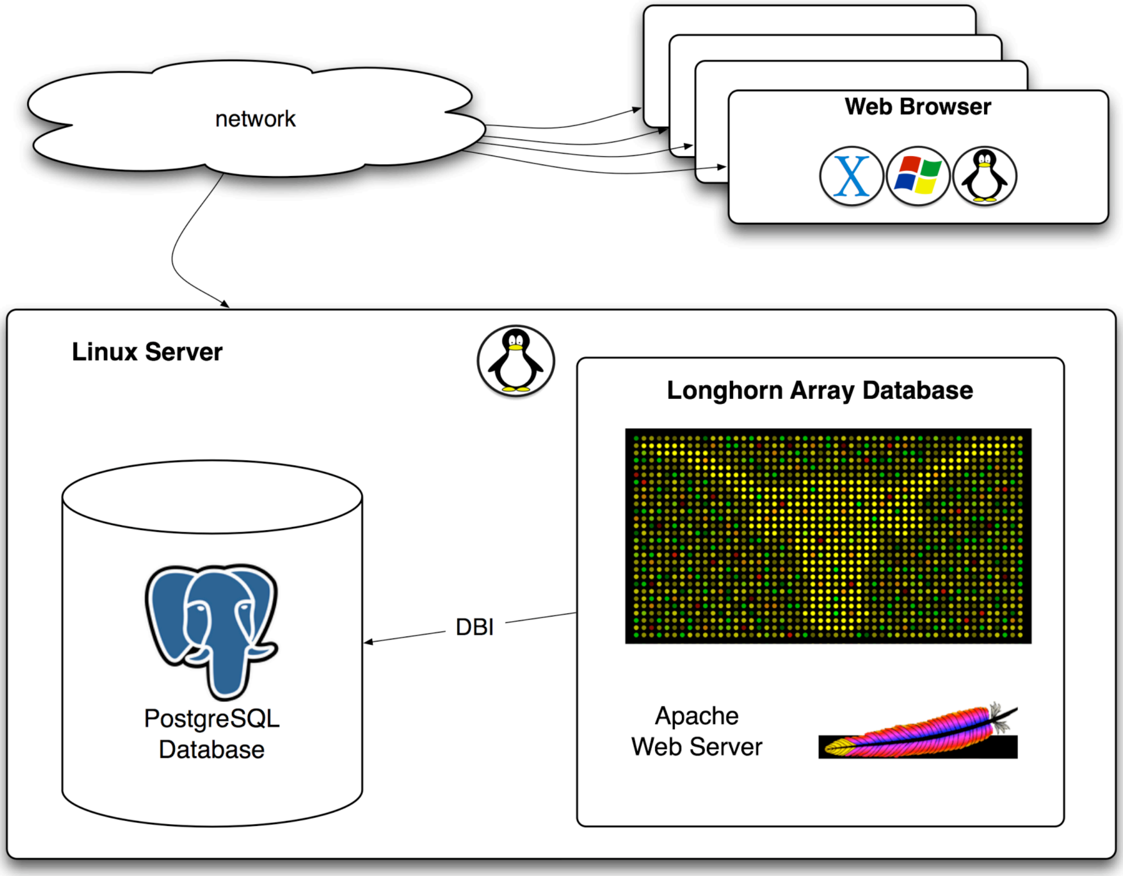
MIAME Support

A critical part of reporting the results of microarray experimentation and data analysis is the ability to share information that fully describes individual experiments. Lack of standards for presenting both experimental conditions and primary microarray data made relative comparison of microarray experiments produced in separate research environments a near impossibility. A standard entitled MIAME – the *Minimum Information About a Microarray Experiment* – was proposed to address this problem [45]. MIAME was a comprehensive specification that detailed the minimum annotation that should accompany the publication of any microarray data set. Subsections pertaining to experimental design, array design, sample preparation, hybridization protocols, actual quantitative results, and normalization controls were all addressed within the specification. MIAME had significant support from both the research and journal publishing communities. It was therefore imperative that any long-term microarray data warehouse and analysis environment support the MIAME specification. This ensured that as results were accumulated within the database the appropriate experimental and conditional annotations were simultaneously recorded and archived.

To enable LAD MIAME compliance we implemented a strategy that allowed a MIAME addendum to be attached to each experiment that was submitted to the database. Subsequent recall of that experiment also recalled all MIAME annotation information. Because a large fraction of this required information remained constant from experiment to experiment, MIAME annotation was implemented through use of reusable templates. This design enabled MIAME compliance without encumbering the experiment submission process.

Figure 3.1 – Longhorn Array Database Core Technology

The Longhorn Array Database relies entirely on open-source technologies. The application component is integrated with the Apache Web Server while PostgreSQL provides the relational database functionality. All LAD functions are web-accessible allowing a distributed user-base to submit and collaboratively share experimental results.



Final Release & Support

Simultaneous to its publication in BMC Bioinformatics, LAD was released to the research community in August of 2003 [40]. Five revisions followed, each including both new features and defect repairs.

Secondary Data Association & Analysis

The development of LAD provided a reliable and secure primary data warehouse. It is difficult to functionally analyze primary microarray data, however, without many forms of exogenous information directly relevant to the set of significant microarray spots. Following raw data retrieval is a process that involves the association of custom and curated annotations, genomic ontology, and raw genomic DNA sequence. These resources are difficult to manually assemble while data manipulation and spot-to-external-data association errors are frequent. Amplifying this problem is the fact that this process is typically not executed just once in the life cycle of a research project. These operations are part of a much larger cycle of analysis requiring execution with each round of hypothesis revision.

To better handle this pervasive set of problems we have developed a network-centric software environment (ArrayPlex) chartered with the goal of streamlining the up-to-date acquisition of these resources and the ease by which they can be associated with primary data. We have done this focusing upon resources related to three model organisms often studied in DNA microarray-based research: *Homo sapiens* (Hs), *Mus musculus* (Mm), and *Saccharomyces cerevisiae* (Sc). Specifically, we have included commonly utilized systematic annotations from the Stanford Genome Database (SGD), NCBI, and the Gene Ontology (GO) Consortium. Additionally, we provide complete access to the latest UCSC-hosted genome sequence. In addition to data resources, we have assembled access to a suite of frequently utilized DNA sequence analysis toolsets. A complete list of managed resources and toolsets are listed in Tables 3.1 and 3.2, respectively.

Our goal was to develop an open-source, easily installed and maintained, robust, network-centric system with which researchers could construct reusable pipelines of complex data analysis procedures. We designed the system to communicate on three levels of interaction: a graphical user interface for interactive data manipulation, a set of command-line analytical modules for script-driven analysis, and a documented Java-based programmatic application programming interface (API). Here we describe the systematic architecture of the ArrayPlex environment and the genomic resources within it. Additionally, we demonstrate how ArrayPlex has been proven in the large-scale analysis of a transcriptional regulatory network.

Core Technology, Design, & Network Operation

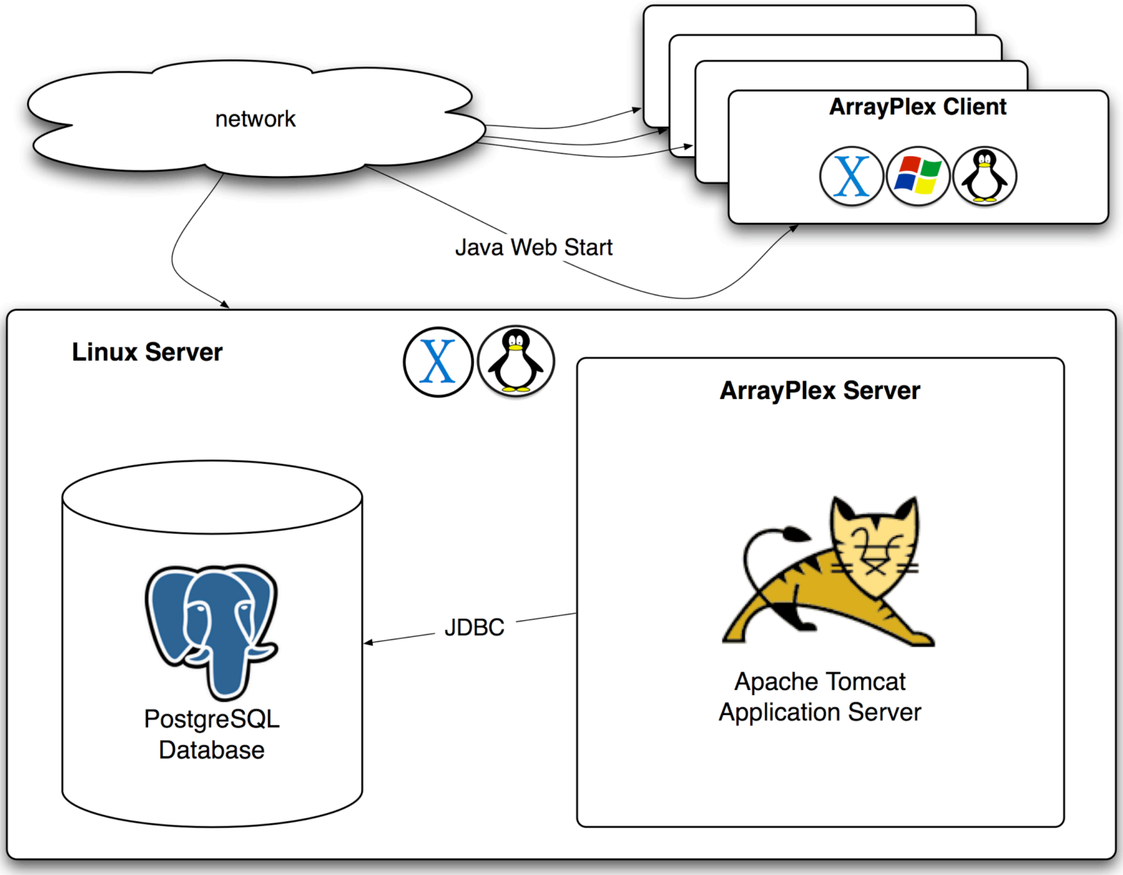
ArrayPlex was implemented with exclusively open-source technologies. Core software components were selected with a set of criteria. The technologies selected were those proven by use in larger software ecosystems that included both research and commercial operation. Components were selected in such a manner that it was possible to create a nearly encapsulated system; a package that when installed was ready to operate without a bevy of additional system requirements to be fulfilled.

The ArrayPlex server is designed to operate on either the Linux operating system or Apple Mac OS X (Figure 3.2). The primary server component of ArrayPlex is the application server Apache Tomcat. The ArrayPlex server stores the large majority of its managed data in the PostgreSQL relational database system. This is the only functional prerequisite, other than the Apple Mac OS X or Linux operating system itself, of the ArrayPlex server that must be obtained and prepared prior to ArrayPlex server installation.

The ArrayPlex client is a graphical user interface that contains dozens of data management, analysis, and visualization features. It is compatible with the Apple Mac OS X, Windows XP, Windows Vista and most distributions of the Linux operating system. It communicates by standard network protocols with the ArrayPlex server and thus can operate on any computer connected to the ArrayPlex server.

Figure 3.2 – ArrayPlex Core Technology

The ArrayPlex server is a nearly encapsulated system comprised of an embedded Java Runtime Environment and Apache Tomcat application server. The ArrayPlex requires one external resource, a PostgreSQL relational database server. The ArrayPlex server operates within the Linux operating system and communicates with the PostgreSQL server by the standard JDBC protocol. The ArrayPlex client can be operated on any Apple Mac OS X, Microsoft Windows, or Linux computer. The ArrayPlex client is not installed but rather launched through use of Java Web Start. This ensures that the ArrayPlex client is always up-to-date when used on any laptop or desktop computer. The ArrayPlex client communicates with the ArrayPlex server by HTTP, the same protocol by which web browsers communicate with web servers. Thus, the ArrayPlex client can be used anywhere there is simple network connectivity back to the ArrayPlex server.



Because it communicates with the ArrayPlex server using the same protocol a web browser utilizes, the ArrayPlex client requires no special changes to client firewall configurations or network settings for operation. The ArrayPlex client requires no local installation process to be run in order to operate. The application resides on the ArrayPlex server and is retrieved and launched through use of Java Web Start. This ensures that with each execution the end-user is using the latest version of the ArrayPlex client. This implementation allows a large research group to share a customizable and expanding graphical user interface without the perpetual need for widespread upgrade or reinstallation with each cycle of improvement. In addition to the graphical user interface, ArrayPlex has a set of command-line executed client-side modules packaged in the form of standard Java Archive format (JAR) files. These modules contain documented analytical routines that use the network to communicate with the ArrayPlex server in the same way that the ArrayPlex client does. This allows the distributed network design of ArrayPlex to be used by command-line application and script-driven analysis as easily as the graphical interface.

Bundled Genomic Resources

The complete ArrayPlex server meta-environment is composed of the combination of the ArrayPlex application server and the many genomic resources and analytical toolsets it provides (Figure 3.2, Table 3.1, and Table 3.2). The process of ArrayPlex server installation acquires each of the genomic resources (Table 3.1) from its officially hosted location. This includes generic GO ontology descriptors, organism-specific GO ontology assignments, and organism-specific gene annotations.

Figure 3.3 – ArrayPlex Architecture, Resources, & Communication

The complete ArrayPlex environment is composed of the combination of the ArrayPlex application server and the many genomic resources and analytical toolsets that it installs, manages, and provides. The ArrayPlex server installs genomic annotations, ontological assignments, and genome sequence. Additionally, toolsets providing genomic sequence extraction, BLAST, sequence search, sequence discovery, and multi-sequence alignment are provided.

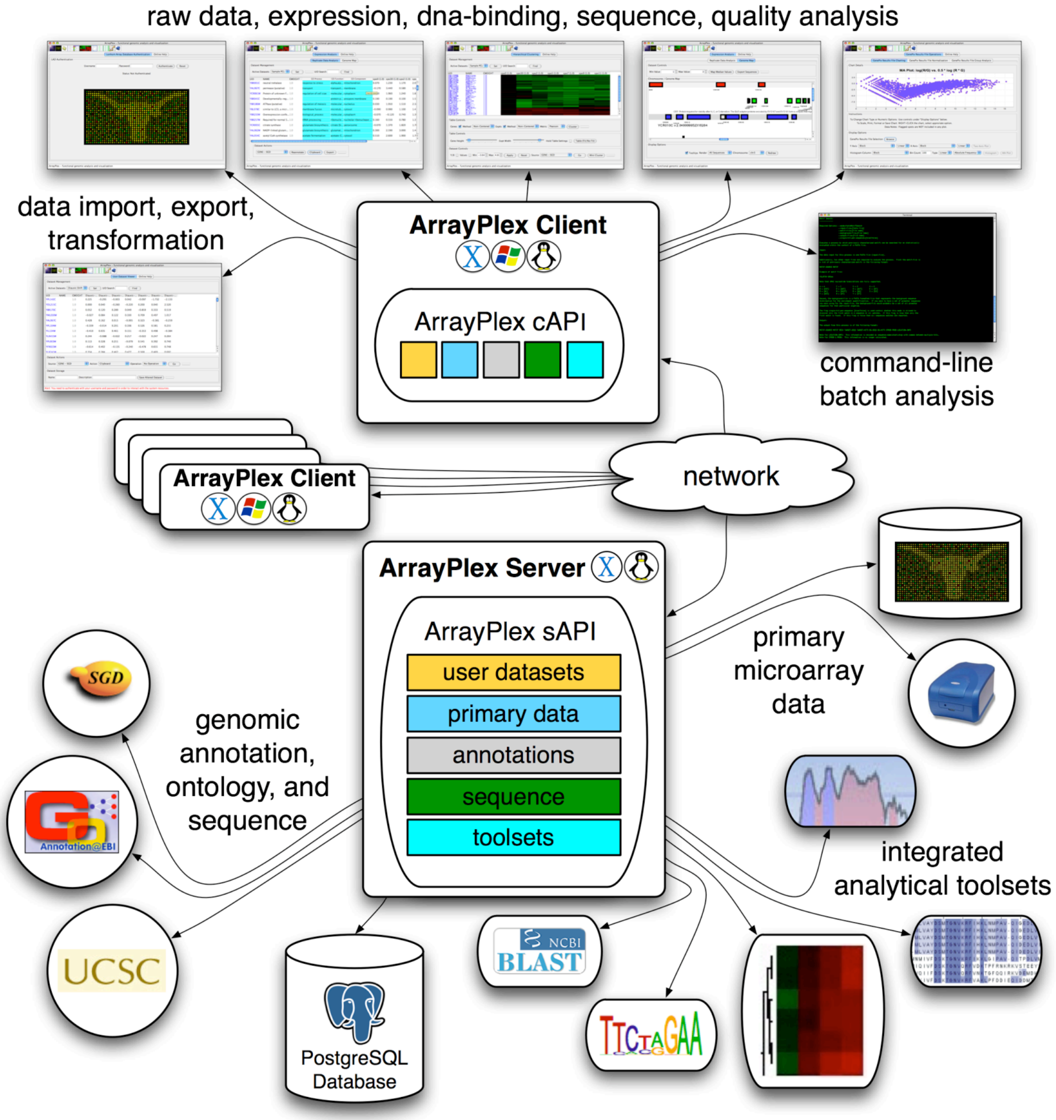


Table 3.1 – ArrayPlex Managed Resources

Genomic resources downloaded by the ArrayPlex installation program. Each of these resources is kept up-to-date and is accessible by the ArrayPlex client, command-line modules, and programmatic API.

Resource Name	Source	Related Organism
GO Ontology Descriptors	GO Consortium	Hs, Mm, Sc
Genome Sequence	UCSC	Hs, Mm
Hs GO Ontology Assignments	EBI	Hs
Mm GO Ontology Assignments	EBI	Mm
Sc Annotations	SGD	Sc
Sc Genome Sequence	SGD	Sc
Sc GO Ontology Assignments	SGD	Sc

Table 3.2 – ArrayPlex Integrated Toolsets

The toolsets integrated into the ArrayPlex server environment. The download code of *Bundle* indicates that that ArrayPlex installation program is capable of downloading the source-code and building the tool during the installation process with no further interaction needed. Alternatively, a code of *Acquire* indicates that a license agreement is required for download and thus the installer of the ArrayPlex server must manually download a file and place it in the proper place on the ArrayPlex server. Documentation is provided for how to acquire and install all toolsets with this requirement.

Tool Name	**Purpose**	**Download**
AlignAce	sequence discovery	Acquire
Avid	sequence alignment	Acquire
BLAST	genomic sequence matching	Bundle
ClustalW	sequence alignment	Bundle
cluster	hierarchical clustering	Acquire
MDSCAN	sequence discovery	Bundle
MEME	sequence discovery	Bundle
fastacmd	sequence retrieval	Bundle
rVista	sequence alignment	Acquire

Table 3.3 – ArrayPlex Modules

The six command-line modules built by and provided with the ArrayPlex installation. The first three modules, classified a *Generic* are the modules most useful to a researcher desiring command-line access to any of the resources hosted on the ArrayPlex server. This includes all genome sequence, annotation, ontology, and user dataset information. The *SequenceAnalysis.jar* module, additionally, contains all of the genome sequence operations featured in the ArrayPlex client including organism-specific sequence extraction, BLAST, known-motif search, motif discovery, and multi-sequence alignment. The modules classified as *Regulation* are generally usable but were developed in a very specialized context. These modules are the nearly-complete computational infrastructure for a recent large-scale study of systematic deletion of 263 individual transcription factors in *Saccharomyces cerevisiae*. They provide both reusable analytical operations and a guide as to how the ArrayPlex programmatic API can be used for constructing novel analysis routines.

Module Name	Purpose	Class
AnnotationResources.jar	Genome annotation and ontology retrieval	Generic
DatasetOperations.jar	User dataset retrieval, transformation, and manipulation	Generic
SequenceAnalysis.jar	Genome sequence extraction, search, discovery, manipulation	Generic
ErrorModel.jar	Example routines in replicate combination	Regulation
InteractionGraph.jar	Example routines in network modeling	Regulation
TargetAnalysis.jar	Example routines in ontological and sequence analysis	Regulation

All resources are processed from their heterogeneous downloaded forms to a structured query language (SQL) format that is loaded into the ArrayPlex relational database schema. The design and implementation of this data transformation was central to the end-goal of the ArrayPlex programmatic API: generic, organism-independent, reusability of core objects and analytical routines. The transformation removes all of the organism-specific nature of the data and allows the ArrayPlex programmatic API to be designed such that reusable code modules can be implemented independent of the original source of the information being transported.

A functional example of this would be GO ontology assignments. This information is species-specific and details the mapping of universal GO ontology terms to specific genes within a given organism. The downloaded content of these assignments for human and mouse differ from yeast in format and content. This is the result of the fact that these assignments are curated and managed by independent research institutions: EBI for human and mouse, SGD for yeast. The transformation of this information to a single format and normalized storage in a relational schema allowed for a single set of ArrayPlex database source-code to be written to retrieve and use this information. This allows programmers using the ArrayPlex programmatic API to write data retrieval and analysis routines that are independent of the organism-specific caveats and institution-specific file formats.

In addition to GO ontology and gene annotations, complete genome sequence is downloaded for each of the supported model organisms. This genome sequence is FASTA in raw form but is converted to NCBI BLAST-database format by the ArrayPlex installation program using NCBI-provided utilities. This transformation is performed for two reasons. First, it allows the ArrayPlex programmatic API to include complete BLAST functionality as a part of its catalogue of analytical operations. Of more importance it allows the ArrayPlex environment to leverage the NCBI-bundled toolsets for genome sequence retrieval. These toolsets are designed to be both rapid and precise in their retrieval and manipulation of genomic sequence. The process of acquiring and understanding the large set of NCBI-provided utilities for these operations can be inefficient. Their bundled-inclusion in ArrayPlex is intended to provide their benefits while simplifying the process by which they are used.

Genome resources are most valuable when synchronized with the most recent versions available. Frequent modifications and additions occur to datasets, especially GO ontology and gene annotation assignments that are curated and updated based on published research. Analytical routines and their biological conclusions suffer when input knowledge such as these assignments are not kept in sync with current revisions. For this reason, the ArrayPlex system is designed to not only download and store this information upon system installation, but also to have the capacity to check for updated information, retrieve it, and update the resources managed within the relational schema. This functionality is provided and documented in the format of a standard system scheduler that is a part of all Linux server environments.

Integrated Sequence Analysis Toolsets

In addition to the many genome resources hosted on the ArrayPlex server a complete set of open-source analytical toolsets are integrated into the environment (Table 3.2). The set of tools include NCBI BLAST, cluster, CLUSTALW, AVID/rVista, and several sequence motif discovery applications: AlignAce, MDSCAN, and MEME. As detailed in Table 3.2, the majority of these applications are downloaded, compiled from source-code, and installed by the ArrayPlex installation program. The limitation of licensing agreements made this not possible for a few of the integrated toolsets.

Complete documentation is included with the ArrayPlex installation on how to retrieve and install these additional utilities. The inclusion of these toolsets transformed ArrayPlex from solely an information warehouse to a server capable of extended analytical capacity. All of these analytical features are accessible by way of the graphical ArrayPlex client application, the command-line modules, and the programmatic API. This access facilitates high-throughput data and sequence operations such as sequence retrieval, data manipulation and transformation, multi-genome BLAST, sequence motif search and discovery, hierarchical clustering, and sequence alignment. The execution of these utilities on the ArrayPlex server provides their analytical functions without any direct support needed by either the ArrayPlex client or command-line modules. In this fashion, it is now possible to retrieve the information provided by these utilities from computers that might never have been able to otherwise compile or run them.

Analytical Accessibility & Customization

In addition to the many genome resources and toolsets hosted by the ArrayPlex environment, Figure 3.2 depicts the overall interactivity and relationship of the subcomponent elements. Both the ArrayPlex client and the command-line modules communicate over a network connection with the ArrayPlex server using the HTTP protocol. It is both possible and intended for many individual clients or command-line modules to simultaneously interact with a single server. Our research process included many multi-week executions of more than a dozen command-line modules interacting with a single ArrayPlex server for annotation, ontology, and genome sequence, as well as analytical toolset executions. The ArrayPlex server was easily able to handle these parallel requests.

The details by which the ArrayPlex programmatic API communicates from either client or command-line module to server are depicted in Figure 3.4. Demonstrated are matching sequence motif analysis features from both the graphical client and the command-line module *SequenceAnalysis.jar* (Figure 3.4). Each of these components interacts with the API by way of the `[net.sourceforge.arrayplex.client]` package of routines. These client routines are designed to marshal the input parameters, data, and named operations being sent to them in such a way that the ArrayPlex server can decode this information and respond. The objects transportable from client to server and back are an extensive and specialized set that is part of the `[net.sourceforge.arrayplex.serial]` package of resources. The `[net.sourceforge.arrayplex.servlet]` package receives requests and decodes both what part of the client API made the request and what specific information is being sent to facilitate it. The servlet API then calls a mirror server API based upon this information. This server API, which is packaged as `[net.sourceforge.arrayplex.server]`, is where actual functional operations begin to take place. This package contains dozens of classes that interact with the ArrayPlex server operating system to execute analytical tasks or with the ArrayPlex relational database API `[net.sourceforge.arrayplex.db]` to retrieve either user datasets or genomic annotations. When either an analytical process completes or database-stored information is retrieved, the process begins to fold back upon itself. Information is again loaded into API-based objects that are returned across the network to the original client operation.

This design is notable in two ways. First, the bioinformaticist utilizing the client API routines needs no actual knowledge that the programmatic request will be fulfilled over a network on a remote server. The API is designed such that the complication of network implementation is hidden from the user. For example, the operation *executeBlastAll (organism, evalue, sequence)* that is part of the documented *SequenceResources* client API, gives no clue to the programmatic user that the implementation of the actual operation requires that the parameters *organism*, *evalue*, and *sequence* be encoded into an object and sent across the network to the ArrayPlex server where the NCBI-BLAST utility *blastall* is executed. The result of that *blastall* execution is then read from the server file-system, formatted into a programmatic object, and returned across the network to the client computer. To the programmatic user of the client API no network operation is either evident or noted; the *BlastResult* object is the result of the operation and their programmatic routines move to the next step. It is as if everything executed and completed on their local computer.

Figure 3.4 – ArrayPlex Client & Server API Network Operation

Both the ArrayPlex client and command-line modules use the network capabilities of the ArrayPlex API to send requests and retrieve results.

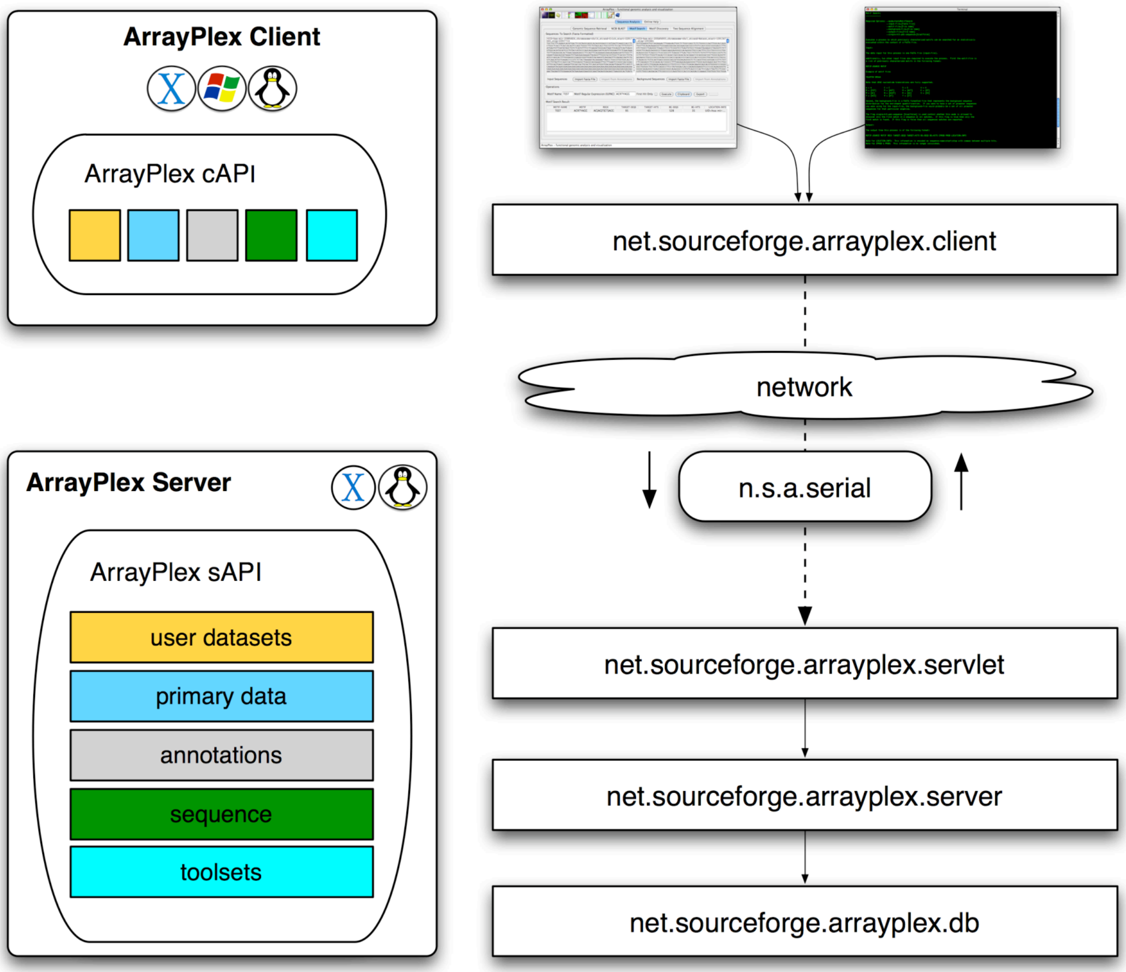
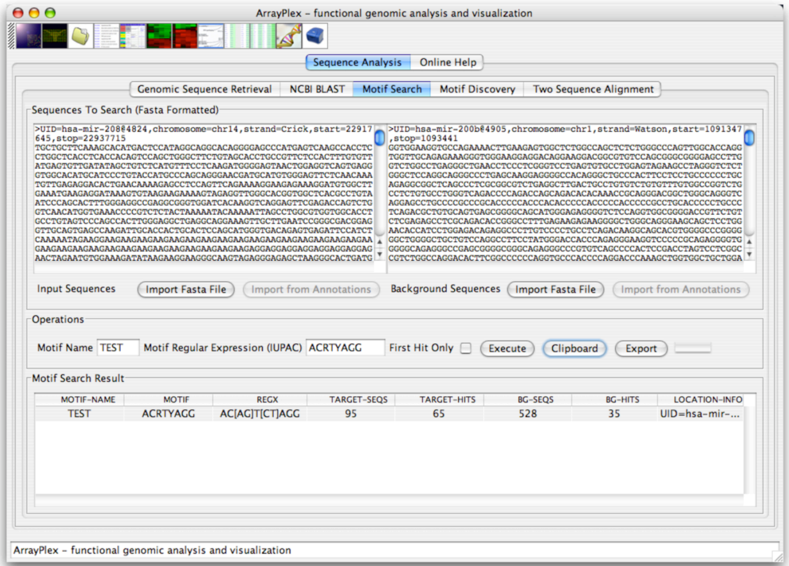


Figure 3.5 – ArrayPlex Client & Module Pairing

The ArrayPlex client and command-line modules have matching sets of operations for nearly all genome resources and analytical toolsets. Pictured are the graphical and command-line module versions of a sequence motif search function. The option for specifying input FASTA sequence, background FASTA sequence, a motif to search for, and reporting of only *first hits* or all hits within a single sequence are provided for within both contexts.



```
>UID=hsa-mir-208@4824,chromosome=chr14,strand=Crick,start=22917
645,stop=22937715
TGCTGCTTCAAAGCACATGACTCCATAGGCAGGCACAGGGGAGCCCATGAGTCAAGCCACCTC
CTGGCTCACCTCACCACAGTCCAGCTGGGCTTCTGTAGCACCTGCCGTTCTCCACTTTGTGTT
ATGAGTGTTGATATAGCTGTCTCATGTTTCCTCAAGATGGGGAGTAACTGGAGGTCAGTGAGG
GTGGCACATGCATCCCTGTACCATGCCCAGCAGGGAACGATGCATGTGGGAGTTCTCAACAAA
TGTTGAGAGGACACTGAACAAAAGAGCCTCCAGTTCAGAAAAGGAAGAGAAAGGATGTGGCTT
GAAATGAAGAGGATAAAGTGTAAGAAGAAAAGTAGAGGTTGGGCACGGTGGCTCACGCCTGTA
ATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCACAAGGTCAGGAGTTCGAGACCAGTCTG
GTCAACATGGTGAAACCCCGTCTCTACTAAAAATACAAAAATTAGCCTGGCGTGGTGGCACCT
GCCTGTAGTCCCAGCCACTTGGGAGGCTGAGGCAGGAAAGTTGCTTGAATCCGGGCGACGGAG
GTTGCAGTGAGCCAAGATTGCACCACTGCACTCCAGCATGGGTGACAGAGTGAGATTCCATCT
CAAAAATAGAAGGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAA
GAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAGGAGGAGGAGGAGGAGGAG
AACTAGAATGTGGAAAGATATAAGAAGGAAGGGCAAGTAGAGGGAGAGCTAAGGGCACTGATG
```

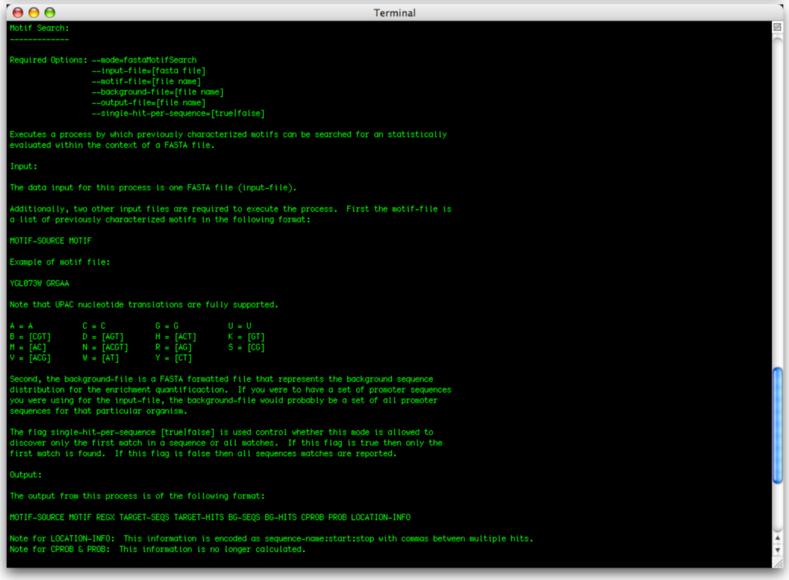
```
>UID=hsa-mir-200b@4905,chromosome=chr1,strand=Watson,start=1091347
,stop=1093441
GGTGGAAGGTGCCAGAAAACTTGAAGAGTGGCTCTGGCCAGCTCTCTGGGCCCAGTTGGCACCAGG
TGGTTGCAGAGAAAGGGTGGGAAGGAGGACAGGAAGGACGGCGTGTCCAGCGGGCGGGGAGCCTTG
GTCTGGCCTGAGGGCTGAACCTCCCTCGGGTCCTGAGTGTGCCTGGAGTAGAAGCCTAGGGTCTCT
GGGCTCCAGGCAGGGCCCTGAGCAAGGAGGGGCCACAGGGCTGCCCACTTCCTCCTGCCCCCCTGC
AGAGGCGGCTCAGCCCTCGCGGCGTCTGAGGCTTGACTGCCTGTGTCTGTGTTTGTGGCCGGTCTG
CCTCTGTGCCTGGGTCAGACCCCAGACCAGCAGACACACAAACCGCAGGGACGGCTGGGCAGGGTC
AGGAGCCTGCCCCGCCCGCACCCCCACCCACACCCCCACCCCCACCCCCGCCTGCACCCCCTGCCC
TCAGACGCTGTGCAGTGAGCGGGGCAGCATGGGAGAGGGGTCTCCAGGTGGCGGGGACCGTTCTGT
CTCGAGAGCCTCGCAGACACCGGGCCTTTGAGAAGAGAAGGGGCTGGGCAGGGAAGCAGCTCCTGG
AACACCATCCTGGAGACAGAGGCCCTTGTCCCCTGCCTCAGACAAGGCAGCACGTGGGGCCCGGGG
GGCTGGGGCTGCTGTCCAGGCCTTCCTATGGGACCACCCAGAGGGAAGGTCCCCCGCAGAGGGGTG
GGGGCAGAGGGCCGAGCGGGGCGGGCAGAGGGCCCGTGTCAGCCCCACTCCGACCTAGTCCTCGGC
CGTCTGGCCAGGACACTTCGGCCCCCCAGGTGCCCACCCCAGGACCCAAAGCTGGTGGCTGCTGGA
```

Motif Name TEST

Motif Regular Expression (IUPAC) ACRTYAGG

First Hit Only ☐

MOTIF-NAME	MOTIF	REGX	TARGET-SEQS	TARGET-HITS	BG-SEQS	BG-HITS	LOCATION-INFO
TEST	ACRTYAGG	AC[AG]T[CT]AGG	95	65	528	35	UID=hsa-mir-...



```
Motif Search:
-------------

Required Options: --mode=fastaMotifSearch
                  --input-file=[fasta file]
                  --motif-file=[file name]
                  --background-file=[file name]
                  --output-file=[file name]
                  --single-hit-per-sequence=[true|false]

Executes a process by which previously characterized motifs can be searched for an statistically
evaluated within the context of a FASTA file.

Input:

The data input for this process is one FASTA file (input-file).

Additionally, two other input files are required to execute the process.  First the motif-file is
a list of previously characterized motifs in the following format:

MOTIF-SOURCE MOTIF

Example of motif file:

YGL073W GRGAA

Note that UPAC nucleotide translations are fully supported.

A = A           C = C           G = G           U = U
B = [CGT]       D = [AGT]       H = [ACT]       K = [GT]
M = [AC]        N = [ACGT]      R = [AG]        S = [CG]
V = [ACG]       W = [AT]        Y = [CT]

Second, the background-file is a FASTA formatted file that represents the background sequence
distribution for the enrichment quantificaction.  If you were to have a set of promoter sequences
you were using for the input-file, the background-file would probably be a set of all promoter
sequences for that particular organism.

The flag single-hit-per-sequence [true|false] is used control whether this mode is allowed to
discover only the first match in a sequence or all matches.  If this flag is true then only the
first match is found.  If this flag is false then all sequences matches are reported.

Output:

The output from this process is of the following format:

MOTIF-SOURCE MOTIF REGX TARGET-SEQS TARGET-HITS BG-SEQS BG-HITS CPROB PROB LOCATION-INFO

Note for LOCATION-INFO:  This information is encoded as sequence-name:start:stop with commas between multiple hits.
Note for CPROB & PROB:  This information is no longer calculated.
```

Second, the information that exchanged with the ArrayPlex server is in the form of documented API objects. This increases the efficiency by which a programmatic user can utilize the ArrayPlex API compared to other methods that launch processes remotely and retrieve results locally. Most methods of remote task invocation require the programmatic user to parse a stream of resulting information that is returned from the server. The task of parsing this information and determining actual results is error-prone. The ArrayPlex APIs are designed to communicate in terms of API documented objects. Using the example above, the *BlastResult* object that is returned from the ArrayPlex server is just that – a programmatic object like any other in the application environment. Referring to the provided documentation the programmatic user can denote that the *BlastResult* object is composed of a set of *BlastHit* objects, each of which have parameters that describe the genomic loci where BLAST found matching sequences.

Finally, the entire ArrayPlex environment is designed to allow customization. The ArrayPlex client can incorporate internationalization and localization of language elements through modification of a single resource bundle containing nearly all labels that appear throughout its interactive graphical interface. Sections of the ArrayPlex client can be removed; newly designed sections can be accommodated.

Documentation and Guidance

Complete use of the ArrayPlex client, command-line-modules, and programmatic API is documented. The command-line execution of the *SequenceAnalysis.jar* module demonstrates the in-line documentation provided by each of the command-line modules (Figure 3.5). Similarly, the ArrayPlex client has hypertext-formatted help content for each of the interactive sections of the application. This content describes the analytical effect of chosen options and the meaning of results that are displayed. The programmatic

API, similarly documented, details the parameters required by each API and both the format and meaning of returned objects.

DISCUSSION

Utility of the Primary Data Warehouse

We initially deployed LAD on an Intel-based dual-Xeon Dell Precision 530 workstation with 1 GB of RAM and 500 GB of hard disk capacity. Simultaneous to the publication of the LAD manuscript, LAD stored approximately 1300 microarray experiments. The rapid influx of experimental data and the expansion of the user base to more than 100 researchers quickly demanded an upgrade in the hardware dedicated to the application. A two-tier application and database architecture was developed that included the dedicated use of an eight-drive multi-terabyte RAID hard-disk enclosure. This environment has been utilized for nearly five years. LAD currently hosts nearly 7300 individual microarray experiments for more than 130 total users. We are currently in the process of moving to the next generation of LAD hardware and software improvements.

LAD Adoption at Research Institutions

The use of both Linux and PostgreSQL greatly reduced the level of complexity required to run a production microarray database due to the ease with which it is both installed and maintained. This opened up the possibility of a larger community of developers becoming involved with a proven array data warehouse. Several research institutions have deployed LAD creating a larger community of both LAD contributors and users.

ArrayPlex Analytical Proving Ground

The entire ArrayPlex system – server resources, client, and all command-line modules, were tested over the course of more than a year in an active research setting. Earlier this year we published the results of an extensive study of 263 individual transcription factor deletions in *Saccharomyces cerevisiae* [46]. The ArrayPlex system was the central hub of all analytical activities for this research. The complete experimental design and results of this research are presented in Chapter 4.

The published manuscript details the results of extensive GO ontological enrichment analysis, sequence motif search, and novel sequence motif discovery. The ArrayPlex command-line modules *ErrorModel.jar*, *InteractionGraph.jar*, and *TargetAnalysis.jar* (Table 3.3) were central in defining the set of operations that led to the resulting biological conclusions. These modules are included as part of the ArrayPlex set of command-line functions as their capacity is useful to any mRNA expression-based microarray research. Additionally, the command-line modules *AnnotationResources.jar*, *DatasetOperations.jar*, and *SequenceAnalysis.jar* provide abstract implementations of methods to expose the genome sequence and resources hosted by the ArrayPlex server to the command-line module user.

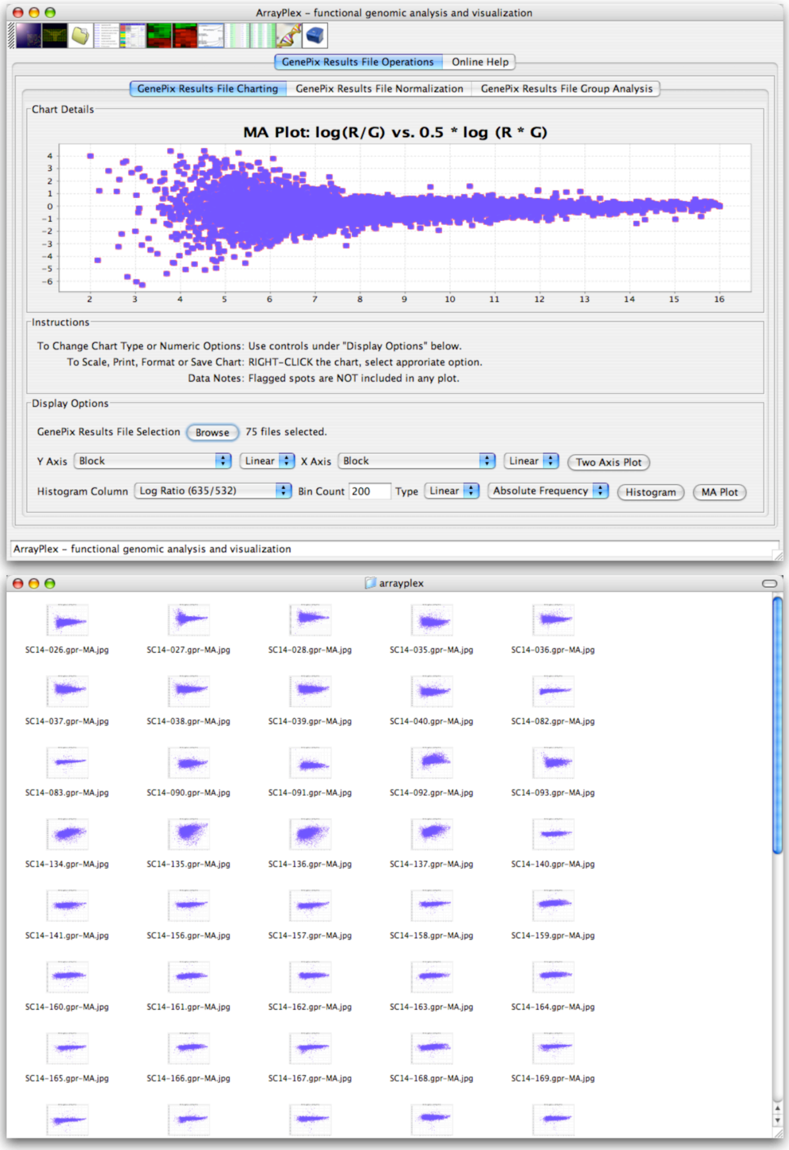
High-throughput Microarray Data Quality Analysis

A significant data processing step that precedes actual functional analysis in DNA microarray research is one of data quality evaluation. It is essential to understand the quality of each microarray experiment, check for any signal bias, and understand the effect that normalization has on individual and grouped batches of microarray experiments. Secondarily, the selection of significant microarray values for an individual or set of experiments involves the filtering of candidate spots based on a variety of spot metrics. Measurements such as signal to noise ratios, spot consistency regression correlations, and background subtracted single-channel intensity values are typical metrics that are used to separate statistically believable spot values from those that might be of spurious quality.

To address these issues we developed an entire section of the ArrayPlex client dedicated to processing, statistical analysis, and visualization of large batches of input data. The *GenePix Results File Operations* section of the ArrayPlex client has the capacity to batch-process an unlimited quantity of Molecular Devices GenePix Pro Result files (GPR) in three ways. First, the *GenePix Results File Charting* section can read large sets of GPR files into a batch queue for graphical analysis. This is useful for the production of batch sets of data bias visualizations such as MA plots, which detect a bias in the relationship of spot absolute signal intensity to spot ratio. Figure 3.6 depicts the selection of seventy-five GPR files and the rapid batch production of seventy-five matching MA plots. These plots were automatically saved to the client file-system and were screened using the image-thumbnail browsing capacity of any operating system file browser.

Figure 3.6 – GenePix Result File Batch Quality Screen

The *GenePix Results File Operations* section of the ArrayPlex client contains extensive resources for the statistical and visual processing of Molecular Devices GenePix Pro Result files (GPR). Batch production of qualitative visualizations such as MA plots, two-axis plots, and spot-metric histograms are possible. This provides the capacity to screen for a number of data quality attributes in large sets of DNA microarray experiments.



To Change Chart Type or Numeric Options: Use controls under "Display Options" below.
To Scale, Print, Format or Save Chart: RIGHT-CLICK the chart, select approriate option.
Data Notes: Flagged spots are NOT included in any plot.

GenePix Results File Selection: 75 files selected.
Y Axis: Block; Linear
X Axis: Block; Linear
Histogram Column: Log Ratio (635/532)
Bin Count: 200
Type: Linear; Absolute Frequency

The graphing capacity of this section of the ArrayPlex client is not limited to MA plots. Both histograms and two-axis plots can be mass-produced for any GPR spot metric. In this manner we were able to screen hundreds of DNA microarray experiments for biased signal-to-ratio relationships, non-normal log-ratio distributions, and substandard signal to noise distributions with the selection of just a few parameters and the browsing of automatically saved images.

Depicted in Appendix AI-21 is the *GenePix Results File Group Analysis* section of the ArrayPlex client. This section of the client was developed to characterize the specific effect individual filters were having on the batch extraction of high-confidence data for the 263 individual transcription factor deletions and experimental replicates. Our primary data was initially stored and analyzed in the Longhorn Array Database (LAD). This database is designed to accommodate batch-retrieval of filtered spot values across a multitude of submitted and normalized experiments. We discovered, however, that we had limited capacity to understand the specific effect individual filters were having on the selection for or elimination of spot values across microarray experiments.

We developed the capacity for the ArrayPlex system to load large quantities of GPR files into named groups. These GPR groups can then be queried with a combinatorial set of spot-metric filters. The output of each query for each experiment in the GPR group is a count of the number of spots that passed the filter criteria and the global normalization coefficient that would be calculated if these remaining spots were used as the complete set of high-confidence spots. This normalization value was then compared back to the normalization coefficient determined for all spots and guided us as to whether spot-metric filters were selecting for non-representative sub-populations of spots within individual or groups of experiments. Secondarily, the interactive manipulation of spot-metric filters allowed us to determine the threshold values that would select for high quality spots yet not select against moderate quality DNA microarray result sets. Finally, once threshold filters were determined we used the export functionality built into the *GenePix Results File Group Analysis* section of the client to retrieve specific spot metrics across large sets of experiments.

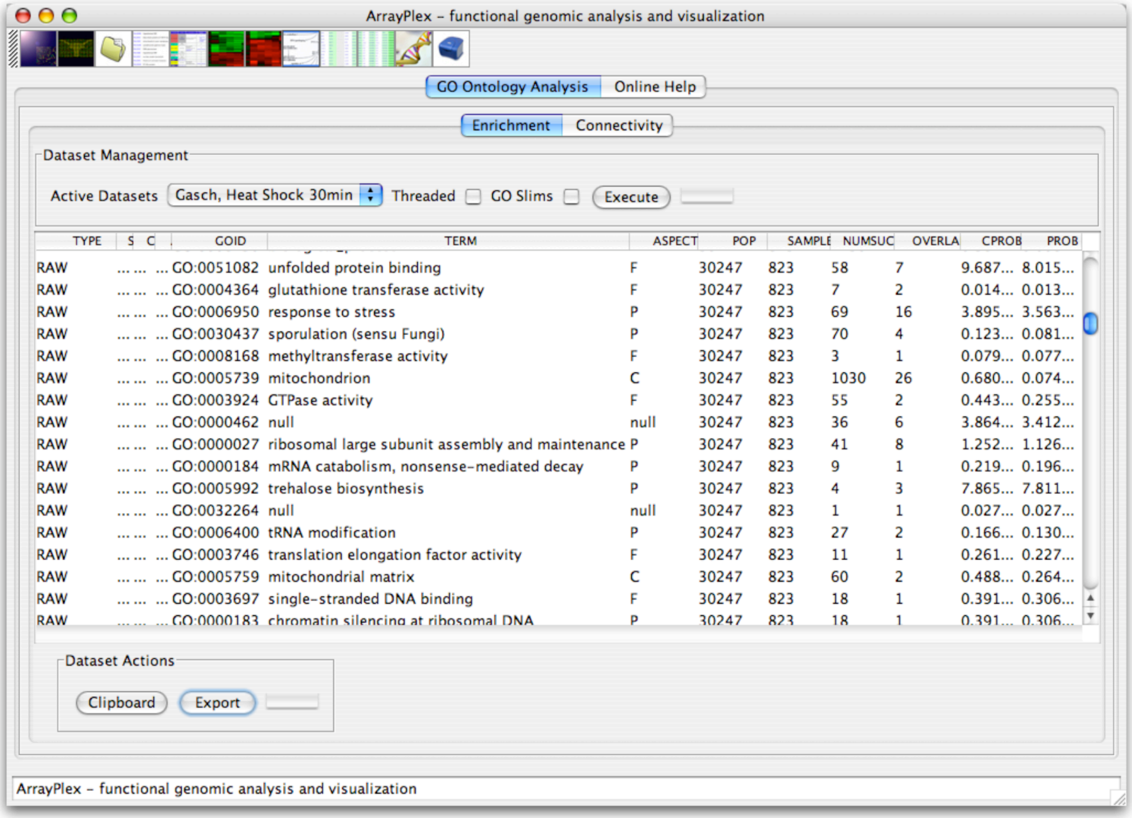
Ontological Enrichment & Connectivity

A successful component of the individual transcription factor deletion data analysis process of was the mining of GO ontological assignments within the pool of target genes affected by a single deletion for GO term-based enrichment. This functionality is built into both the ArrayPlex client and the command-line module *TargetAnalysis.jar*. To cross-reference this functionality we have spot-checked several highly utilized datasets for expected GO term enrichment detection. Figure 3.7 depicts ArrayPlex client processing a *heat-shock* experiment from a published set of environmental stress response data.

Filtering of the GO term enrichment analysis showed that thirty-six terms were significantly enriched at a *P value* less than 0.001 as determined by the cumulative hypergeometric probability distribution. Of these thirty-six terms, a large proportion was shown to specifically relate to *ribosome biogenesis*, *protein folding*, *unfolded protein binding*, *response to stress*, and *chaperone binding*. This is the expected result when mining for GO ontological terms enriched in an experiment where yeast cell cultures were exposed to heat stress.

Figure 3.7 – Ontological Assignment Enrichment & Connectivity

The ArrayPlex client displaying the result of GO term enrichment analysis on previously published environmental stress primary data. Specific GO terms relating to the characterized stress response are clearly enriched.



Active Datasets: Gasch, Heat Shock 30min
Threaded ☐
GO Slims ☐

TYPE	S	C		GOID	TERM	ASPECT	POP	SAMPLE	NUMSUC	OVERLA	CPROB	PROB
RAW	...	...	...	GO:0051082	unfolded protein binding	F	30247	823	58	7	9.687...	8.015...
RAW	...	...	...	GO:0004364	glutathione transferase activity	F	30247	823	7	2	0.014...	0.013...
RAW	...	...	...	GO:0006950	response to stress	P	30247	823	69	16	3.895...	3.563...
RAW	...	...	...	GO:0030437	sporulation (sensu Fungi)	P	30247	823	70	4	0.123...	0.081...
RAW	...	...	...	GO:0008168	methyltransferase activity	F	30247	823	3	1	0.079...	0.077...
RAW	...	...	...	GO:0005739	mitochondrion	C	30247	823	1030	26	0.680...	0.074...
RAW	...	...	...	GO:0003924	GTPase activity	F	30247	823	55	2	0.443...	0.255...
RAW	...	...	...	GO:0000462	null	null	30247	823	36	6	3.864...	3.412...
RAW	...	...	...	GO:0000027	ribosomal large subunit assembly and maintenance	P	30247	823	41	8	1.252...	1.126...
RAW	...	...	...	GO:0000184	mRNA catabolism, nonsense-mediated decay	P	30247	823	9	1	0.219...	0.196...
RAW	...	...	...	GO:0005992	trehalose biosynthesis	P	30247	823	4	3	7.865...	7.811...
RAW	...	...	...	GO:0032264	null	null	30247	823	1	1	0.027...	0.027...
RAW	...	...	...	GO:0006400	tRNA modification	P	30247	823	27	2	0.166...	0.130...
RAW	...	...	...	GO:0003746	translation elongation factor activity	F	30247	823	11	1	0.261...	0.227...
RAW	...	...	...	GO:0005759	mitochondrial matrix	C	30247	823	60	2	0.488...	0.264...
RAW	...	...	...	GO:0003697	single-stranded DNA binding	F	30247	823	18	1	0.391...	0.306...
RAW	...	...	...	GO:0000183	chromatin silencing at ribosomal DNA	P	30247	823	18	1	0.391...	0.306...

ArrayPlex – functional genomic analysis and visualization

Visualization Capacity

The ArrayPlex client and command-line module *InteractionGraph.jar* have the capacity to cross-convert between many commonly used primary data formats. Specifically included is the pre-clustering format (PCL) common to many DNA microarray analysis applications and the graph-markup language format (GML) common to many network-visualization packages such as Cytoscape.

Comparison to Similar Software Packages

ArrayPlex was developed to fulfill the need for interactive, command-line, and programmatic access to up-to-date genomic resources and analytical toolsets in a networked computational environment. Several other research projects have engaged subcomponents of these goals in a variety of ways. EnsMart, Atlas, Mayday, SeqHound, and DAVID are all examples of bioinformatic server environments that address many of the stated associative and analytical goals. SeqHound and Atlas each house an extensive API-accessible list of resources yet lack both an extensible user interface and pre-defined command-line modules. EnsMart has a web interface and command-shell environment but lacks a client-server enabled API. This feature was core to ArrayPlex's design goal of enabling all computers in a research environment to be productive platforms on which data analysis can be accomplished. Mayday and DAVID are toolsets focused upon DNA microarray data analysis and GO ontology analysis, respectively. They each are feature-rich in these categories but lack integration with the wide variety of genomic resources provided by the ArrayPlex environment.

MATERIALS AND METHODS

Longhorn Array Database Requirements & Availability

LAD source-code is freely available to all interested users. The installation manual details the system prerequisites that are required for successful installation and operation of the LAD server. Apache versions 1.x and 2.x are supported. PostgreSQL versions 7.3 and 7.4 have been tested. It is believed that PostgreSQL versions 8.0 and greater will operate but this configuration has not been tested in a production setting.

ArrayPlex Requirements & Availability

ArrayPlex is available from its project site at sourceforge.net. The ArrayPlex server, client, and command-line modules are included in a single installation package. The ArrayPlex client and the command-line modules are prepared during the process of ArrayPlex server installation such that they are configured to communicate with the ArrayPlex server being installed by the system administrator. Complete source-code is provided for each of the operational components.

ArrayPlex Server Requirements

The default server installation requires either an Intel-based computer running the Linux operating system or any computer running Apple Mac OS X. Linux servers running both the 2.4 and 2.6 generation of kernels have been tested and are supported. During its development period, ArrayPlex was operated on Mandrake, Mandriva, Fedora, Gentoo, RedHat, and Ubuntu distributions of Linux. Apple Mac OS X has been tested with version 10.4 (Tiger), but it is believed that most generations of the operating system will be compatible. Additionally, an operational PostgreSQL relational database system is required. The ArrayPlex development and testing process has utilized PostgreSQL server versions 7.3, 7.4, 8.0, 8.1, and 8.2. The database server does not need to be installed on the same computer as the ArrayPlex server, only reachable by TCP/IP network connectivity and standard PostgreSQL client utilities. A sequestered ArrayPlex schema instance is created within the PostgreSQL database server such that ArrayPlex can co-exist with other database instances in operation. Neither a Java Runtime Environment nor an installation of Apache Tomcat is required. Each of these resources is bundled within the ArrayPlex installation in order to create a more encapsulated and ready-to-operate system. Alternative implementations of the Java Runtime Environment or Apache Tomcat can be substituted through simple sub-folder replacement within the installed ArrayPlex server. This process is documented in the *ArrayPlex Server Installation Guide*.

The ArrayPlex distribution, as downloaded from the SourceForge.net project site, is 350MB in size. The ArrayPlex server, however, downloads a large quantity of genomic annotation and sequence during the installation process. The genomic sequence files are transformed into NCBI BLAST-compatible databases that allow for rapid sequence retrieval. This results in the consumption of significant drive space such that an operational ArrayPlex server requires at least 14GB for complete installation.

ArrayPlex Client Requirements

The ArrayPlex client is not installed but rather launched from the ArrayPlex server by clicking a link within any web browser. The client is supported on Apple Mac OS X, Microsoft Windows XP, Microsoft Windows Vista, and most distributions of the Linux operating system. Each of these client operating systems must have a Java Runtime Environment (JRE) installed. The default Microsoft-provided Java installation on any version of Microsoft Windows is not supported. A JRE should be downloaded and installed from Sun Microsystems. The JRE that is bundled with Apple Mac OS X (10.2 Jaguar, 10.3 Panther, 10.4 Tiger, and 10.5 Leopard) has been tested for compatibility.

ArrayPlex Command-Line Module Requirements

The requirements for use of the command-line modules match those of the ArrayPlex client. They are built by the ArrayPlex server installation process and downloaded by a web-browser to any supported client computer.

SUPPLEMENTAL INFORMATION

ArrayPlex Client Feature Set

The complete feature set of the graphical ArrayPlex Client application is presented in Appendix I (1-20).

Chapter 4: A Functional Transcriptional Regulatory Network

The research described in this chapter was performed in collaboration with Dr. Zhanzhi (Mike) Hu. Dr. Hu performed the large-scale expression profiling of more than two hundred individual transcription factor deletions in *Saccharomyces cerevisiae* concomitant to my implementation of the ArrayPlex analytical environment described in Chapter 3. We worked together for over two years to complete the analysis, functional validation experiments, and manuscript. We co-first authored the final manuscript for *Nature Genetics* in April of 2007 (*Genetic reconstruction of a functional transcriptional regulatory network*) [46].

INTRODUCTION

Several research efforts have explored the regulatory relationships of *Saccharomyces cerevisiae* transcription factors to gene targets [47, 48]. These publications have primarily focused upon the measurement and analysis of whole-genome DNA-protein interactions through the capture of transcription factor to genomic DNA binding events. While knowledge of genomic locations in which a transcription factor is shown to bind is supportive to the claim of regulation, it is not conclusive with respect to the true regulatory effect that a transcription factor may or may not be having on a proximal gene target. Several studies have demonstrated that the binding of a transcription factor to the promoter region of a gene does not necessarily result in the activation or repression of that putative gene target. Methods that focus upon DNA-protein interactions of transcription factors are not able to determine whether a true regulatory event has been measured. Additionally, they are unable to measure neither the nature of the regulatory interaction nor the relative strength of the regulatory effect.

Using DNA microarrays, we profiled the whole-genome transcriptional responses of the individual deletion of 263 *Saccharomyces cerevisiae* transcription factors. This experimental approach provided a strategy by which true regulatory targets could be determined for a large set of transcription factors by measuring the actual transcriptional response of regulated genes. The method employed intrinsically captures the active or repressive nature of the regulatory relationship between each factor and its KO-regulated targets as well as the relative strength of the regulatory effect. Additionally, six essential transcription factors were profiled through the use of conditionally repressive *tet-off* transcription factor strains.

Given these experimental qualities and the fact that the scope of experimentation covers nearly all *Saccharomyces cerevisiae* transcription factors, we were able to both uncover the regulatory roles of individual factors profiled and describe a global regulatory network that encompasses all profiled transcription factors, their regulatory interrelationships, and shared target sets.

RESULTS

Primary Data Processing

The expression profiling microarray experiments were performed in duplicate over the course of several months. Data analysis began with processing of the primary microarray data screening for quality, signal bias, and normalization of replicate experiment data into comparable high-confidence target sets.

Bias Screen

As demonstrated in Figure 3.6, the complete set of more than five hundred replicate microarray experiments were individually screened for signal bias through the batch-production of MA plots and log-ratio histograms. The signal bias detected by MA plots is an unexpected relationship between the log-ratio of a microarray spot and its absolute intensity on either of the measured channels. A trend, for example, for high intensity spots to have predominantly high log ratios would be suspect with regards to the putative targets that emerged from such an experiment. The results of the screen showed no significant impact of signal bias on any of the replicate microarray experiment primary datasets.

Strain, Growth & Microarray Normalization

Supplemental Table 4.1 of this chapter provides a complete list of the transcription factors profiled in this study. Additionally, the *Materials and Methods* section provides strain details, growth conditions, and batch-to-batch normalization methods used throughout the execution of the microarray experiments.

Secondary Data Determination by Error Model

DNA microarray experiments are performed in biological replicate. Repeated experimental results are statistically combined in order to determine the set of results that are significant and reliably repeated across replicates. Simple statistical aggregation techniques such as arithmetic mean and median calculations provide a mechanism by which the typical value for a spot can be determined. These methods lack the capacity to communicate the true reliability of the set of measurements across replicates and thus lack the capacity to *weight* final microarray spot values relative to their statistical precision. Additionally, the use of simple data aggregation models requires arbitrary cutoffs to be set in the final data extraction and filtering process. This concept is visualized and described in Figure 4.1. The dark blue and dark red lines represent arbitrary absolute signal intensity and log-ratio cutoff values, respectively. Signal intensity cutoffs are necessary in order to determine a final result set with believable log-ratio values. Low signal intensity on either of the experimental channels can produce near-stochastic variability in log-ratio determination as the reliability of the ratio calculation across channels becomes unstable. Log-ratio cutoffs are used to separate differentially expressed results from those that did not show significant change from transcription factor deletion to wildtype. These methods are insensitive to the log-ratio and signal intensity distributions both within a single experiment and across biological replicates. The arbitrary determination and use of these filters often results in high false positive and false negative rates with regards to spot inclusion and exclusion rates.

We employed an adapted error model in order to determine the final microarray scores across replicate transcription factor deletion experiments [47, 49]. Execution of this method began with the determination of systematic variation that was intrinsic to the DNA microarray experimental process itself. This variation was not simply a byproduct of hybridization but rather was the aggregate variation that was accumulated across the entire process of RNA isolation, reverse transcription, dye incorporation, hybridization, and primary data capture. Multiple iterations of identical RNA labeling and co-hybridization (same vs. same) experiments were performed to produce a set of replicate data that describes the variation within the assay.

The result of these control co-hybridizations was used to produce the *f* statistic such that:

$$f = stdev(\ln(\left(\frac{R}{G}\right)))$$

R, G = the Cy5 and Cy3 channel intensities

For all other experimental data the significance and *P Value* of a measured spot ratio was:

$$X = \frac{(R - G)}{\sqrt{\sigma_1^2 + \sigma_2^2 + f^2 + (R^2 + G^2)}}$$

σ_1^2 = the standard deviation of background Cy5 measurements

σ_2^2 = the standard deviation of background Cy3 measurements

The confidence of any log ratio within a single experiment was expressed as:

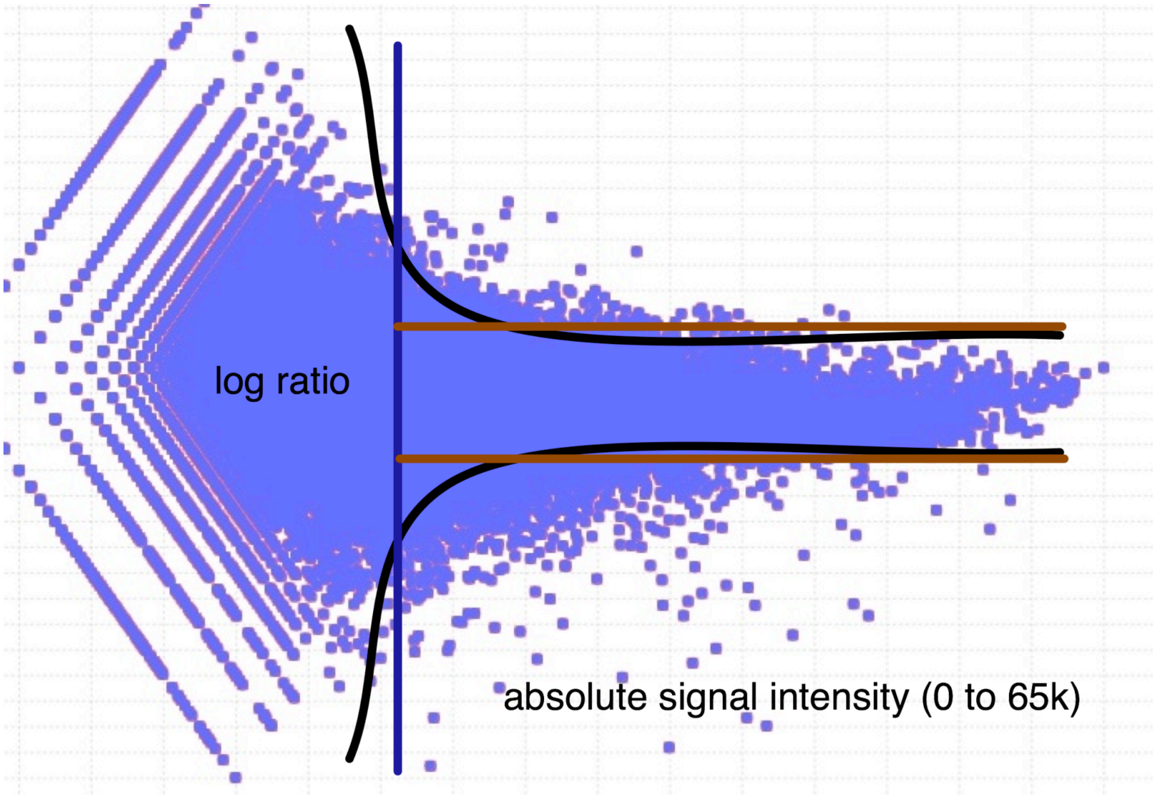
$$\sigma = \frac{\log_{10}\left(\frac{R}{G}\right)}{X}, \; w_i = \frac{1}{\sigma_i^2}$$

The metric w_i was described as the *weight* of any spot within a single microarray experiment. This metric would select against spots with low signal intensity, large errors, and general unreliability by weighting against small *X* values. These calculations provided the metrics needed to determine the statistical precision and significance of all spots in a single microarray experiment. A final metric was needed to combine calculations across biological replicates in such a way that rewarded for repeatability of measurement.

$$(X)score = \frac{\sum_{i=1,n} w_i X_i}{\sum_{i=1,n} w_i}, \; P\ Value = 2 \times normcdf(-|(X)score|)$$

Figure 4.1 – Error Model vs. Arbitrary Filter Cutoffs

The implementation of the error model allowed for a continuous function to be used to determine the significance of all spot-values across biological replicates. The dark blue and dark red lines represent arbitrary absolute signal intensity and log-ratio cutoff values, respectively. These cutoffs are insensitive to the actual foreground, background, and log-ratios present on a single or set of microarray experiments. Additionally, these cutoffs do not have the capacity to select for precise measurements across biological replicates. Alternatively, the error model determines a continuous function (black lines) that incorporates knowledge of the nominal error of the assay as well as the standard deviation of foreground and background measurements for both spots within a single microarray experiment and repeated measurements across replicate experiments.



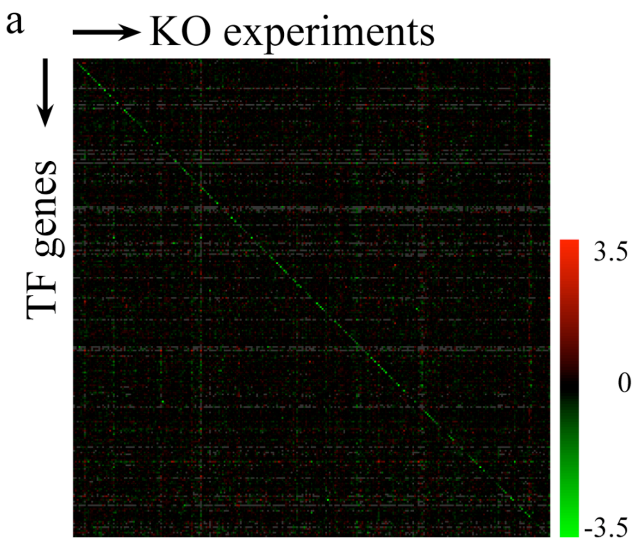
Deletion Expression Validation

We wished to gauge the validity of the deletion strains and corresponding microarray experiments by visualizing the impact of each transcription factor deletion on its own measured transcriptional abundance. The expectation was that measured relative RNA level of the deleted transcription factor should be both in the set of KO-regulated targets and significantly repressed across all transcription factor deletion experiments.

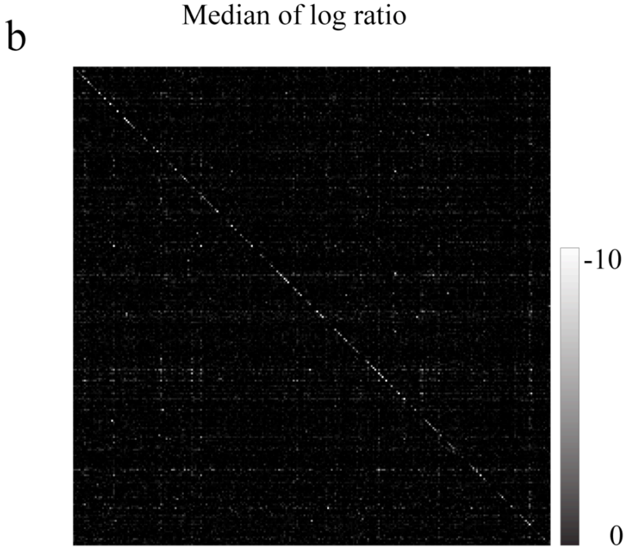
The results of this analysis are presented in Figure 4.1. The data table is sorted along both axes by the transcription factor names of all 263 non-essential transcription factors profiled. Figure 4.1 (a) is rendered by log-ratio. The presence of green and red cells in any single column represents transcription factors that were significantly regulated by the deletion experiment defined by the column in which they appear. The green diagonal in this figure, an intended byproduct of the matching sort along both axes of the plot, confirms the repression of transcription factor genes in corresponding deletion experiments. Figure 4.1 (b) is the same visualization substituting error-model determined *P Values* for log-ratios. The combination of these two visualizations is confirmation that the process of KO-regulated target set determination by the error model produced a set of targets in which the expression of the transcription factor gene was repressed and statistically significant.

Figure 4.2 – Deletion Validation

Each axis contains 263 measurements, ordered by transcription factor name. The appearance of a strong green diagonal in diagram (a) demonstrates the measured transcriptional loss of the transcription factor from its own deletion profile. Diagram (b) uses the *P Value* determined by the error model to demonstrate that the measured loss is both present and significant for nearly all transcription factors profiled.



Median of log ratio



Log *P*-value

Correlative Target Overlap

For each transcription factor profiled we compared the set of KO-regulated targets to targets determined by previous studies [47]. The overlap between our results and the ChIP-chip DNA-protein binding data provided by these studies was low. Comparison of our results, however, with several focused DNA-protein studies of specific transcription factors showed much more favorable overlap [50]. This study determined 354 Rap1 DNA-protein binding target genes. Our results showed 537 KO-regulated targets for this transcription factor. The overlap of 144 common targets was significantly more than the 71 targets that were shared between our result set and the high-throughput ChIP-chip data first used for comparison.

The Functional Regulatory Network

The dataset produced by more than two hundred individual transcription factor deletions was uniquely capable of describing a true transcriptional regulatory network. Each transcription factor deletion had a set of KO-regulated targets that were evaluated to be significant by the process of error-model analysis. Within any set of these KO-regulated targets there were both non-regulatory genes and transcription factors. This phenomenon was expected. It has long been known that the transcriptional abundance of any regulator might be the product of a regulatory cascade; a set of regulator-to-regulator interactions that produces a tuned cellular effect. With many of the transcription factor target pools containing KO-regulated targets that were indeed transcription factors themselves, we proceeded to aggregate the complete set of regulators and targets into a complete functional transcriptional regulatory network.

Tertiary Data Model

Computational analysis of the complete set of transcription factors and KO-regulated targets required a robust system in which the relationships between these items could be accurately and informatively modeled. The construct we chose was the directed weighted graph. A graph (*G*) is a set of objects called vertices (*V*) connected by links denoted as edges (*E*). A directed weighted graph is a graph in which edges between vertices have both an intrinsic directionality and an associated numeric weight. In graph theory, directed weighted graphs are often simply referred to as networks. We mathematically rendered graphs such that vertices represented either transcription factors or KO-regulated targets. Any single edge $e = \{v1, v2\}$ between two vertices described the directional relationship of transcription factor to a single KO-regulated target. The associated weight could take one of two quantitative metrics. Most often it represented the statistical significance of the regulatory relationship (the *P Value*). Alternatively the weighted values reflected the (X) significance score.

The aggregation of all edges emanating from a single transcription factor vertex defined the complete set of KO-regulated targets for that particular transcription factor. This produced a transcription factor-specific sub-network KO^n. The edge set for a factor-specific sub-network was defined as $E(KO^n)$ while the set of vertices were defined as $V(KO^n)$. Co-integration of each sub-network produced a final network that communicates both the specific KO-regulated targets of individual transcription factors and the transitive interrelationships between all factors and the union of their KO-regulated targets.

Final integration provides a declaration of the formal edge set of the network N: $E(N)$ is the set of all unique edges across the union of each $E(KO^n)$ while $V(N)$ (the set of all vertices within the network) is the set of all unique vertices across the union of each $V(KO^n)$.

The relationship between each sub-network KO^n and the final network N:

$V(N)=\bigcup_{i=1}^{n} V(KO^n)$	yields→	$V\left(KO^n\right)\subseteq V\left(N\right)$	yields→	$G\left(KO^n\right)\subseteq G\left(N\right)$
$E(N)=\bigcup_{i=1}^{n} E(KO^n)$	yields→	$E\left(KO^n\right)\subseteq E\left(N\right)$		

The input to all analytical processes was in the form of an *interaction file*. An interaction file is a tab-delimited three-column text file non-sequentially listing transcription factor (A) to KO-regulated target (B) relationships as "A_α B_β Z_β" where α was the set of all transcription factors (269) and β was the set of all KO-regulated targets for a specific transcription factor in α. The weight Z_β was always specific to the interaction between one transcription factor and a single KO-regulated target.

An example of interaction file format is described in Figure 4.3. In this example two transcription factors A and Z are regulating six and nine KO-regulated targets, respectively. Figures 4.4 and 4.5 demonstrate how the individual sub-networks represented by each of these transcription factor KO-regulated target sets can be created and subsequently combined into a singular network containing both factors and the union of their target sets. The X score network implementation of Figure 4.5 is able to communicate both interaction between transcription factors and targets and the nature of the regulatory relationship between them. Interaction files that contained *P Values* and (X) scores were often used simultaneously when both sets of communicated information were needed to form analytical conclusions.

Figure 4.3 – Sample Interaction File

The default format of all KO-regulated target-set data analysis. In this example two transcription factors A and Z are regulating six and nine targets respectively. The left interaction file contains numeric weights that represent the statistical significance between factor and KO-regulated target. Alternatively, the right interaction file uses the weighted X score to communicate the nature of the regulatory relationship (activation or repression). Interaction files containing *P Value* and X score information were often used together when both sets of information were required by the analysis performed.

P Value			**Weighted X Score**		
A	B	2.37E-10	A	B	6.32
A	C	1.23E-08	A	C	-5.33
A	D	1.49E-08	A	D	-5.23
A	E	2.24E-07	A	E	4.80
A	F	3.19E-07	A	F	4.73
A	G	6.06E-12	A	G	7.22
Z	A	8.66E-14	Z	A	8.53
Z	B	9.86E-07	Z	B	4.32
Z	C	1.84E-13	Z	C	8.33
Z	G	2.58E-13	Z	G	8.21
Z	H	2.02E-12	Z	H	7.99
Z	I	2.67E-12	Z	I	7.91
Z	J	2.87E-12	Z	J	7.87
Z	K	3.05E-12	Z	K	-7.63
Z	L	3.40E-12	Z	L	-7.53

$V(KO^{A}) = (\{A, B, C, D, E, F, G\})$

$E(KO^{A}) = (\{A, B\}\ \{A,C\}\ \{A,D\}\ \{A,E\}\ \{A,F\}\ \{A,G\})$

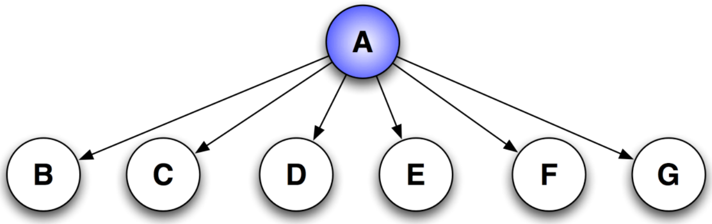
$V(KO^{Z}) = (\{Z, A, B, C, G, H, I, J, K, L\})$

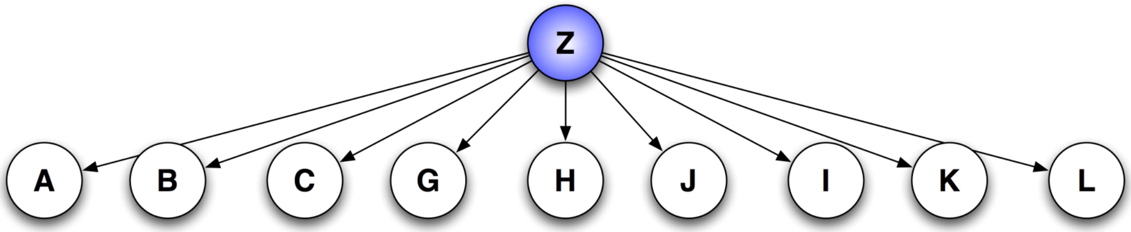
$E(KO^{Z}) = (\{Z, A\}\ \{Z,B\}\ \{Z,C\}\ \{Z,G\}\ \{Z,H\}\ \{Z,I\}\ \{Z,J\}\ \{Z,K\}\ \{Z,L\})$

Figure 4.4 – Network Modeling by *P Value*

The *P Value* interaction file yields the two *sub-networks* $G(KO^A)$ and $G(KO^B)$:

$V(N)$ = ({A, B, C, D, E, F, G, H, I, J, K, L})
$E(N)$ = ({A, B} {A,C} {A,D} {A,E} {A,F} {A,G} {Z, A} {Z,B} {Z,C} {Z,G} {Z,H} {Z,I} {Z,J} {Z,K} {Z,L})





The union of these *sub-networks* $G(KO^A)$ and $G(KO^B)$ yields a final network $G(N)$:

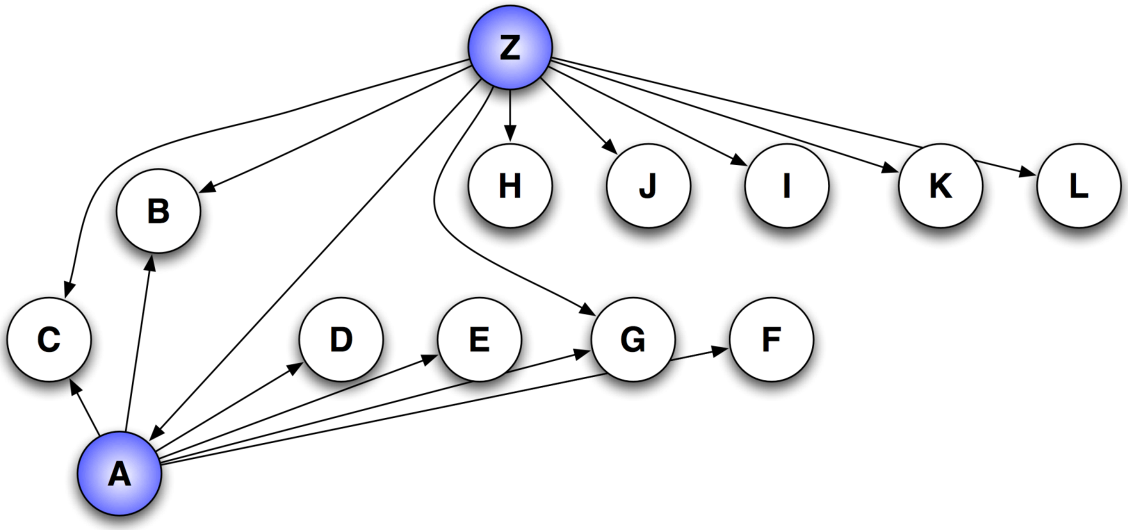
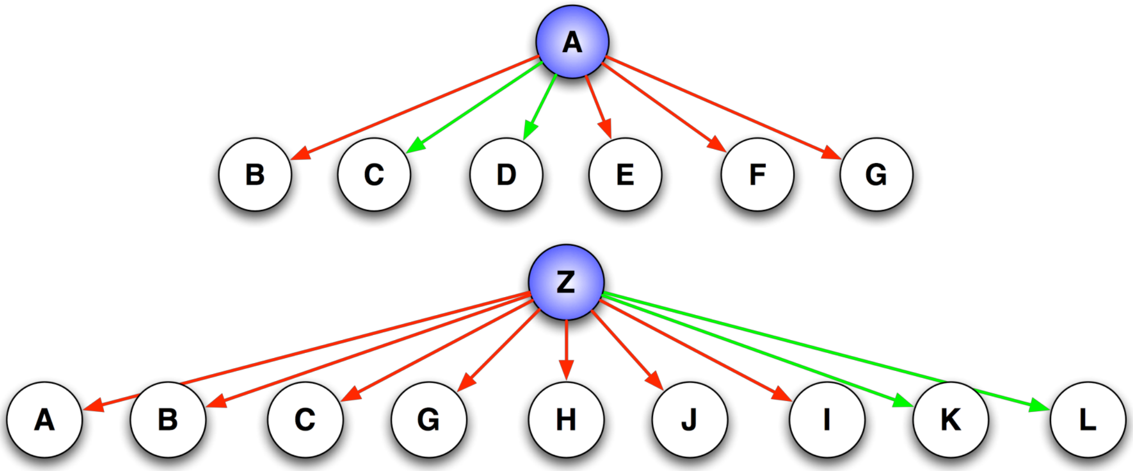


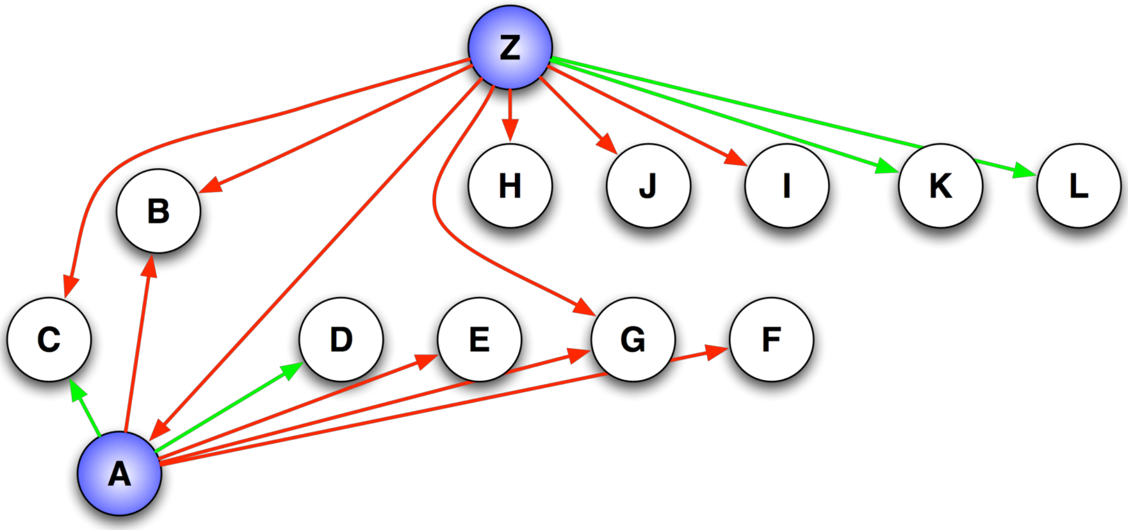
Figure 4.5 – Network Modeling by Weighted X Score

The weighted X score interaction file yields the two *sub-networks* $G(KO^A)$ and $G(KO^B)$:

$V(N)$ = ({A, B, C, D, E, F, G, H, I, J, K, L})
$E(N)$ = ({A, B} {A,C} {A,D} {A,E} {A,F} {A,G} {Z, A} {Z,B} {Z,C} {Z,G} {Z,H} {Z,I} {Z,J} {Z,K} {Z,L})



The union of these *sub-networks* $G(KO^A)$ and $G(KO^B)$ yields a final network $G(N)$:



The Regulatory Network G(N)

The combined network $G(N)$ carried significant regulatory information. Each transcription factor was connected to its pool of KO-regulated targets through directed, weighted edges that described both directionality of regulation and relative significance. All vertices that were the origin of directed vertex-to-vertex connections were by definition known to be transcription factors as only transcription factors had the capacity to be the origin of a *directed* network edge. All vertices that were targets of a directed edge could be either transcription factors or non-regulatory gene targets.

Several key issues were central to the study with regards to understanding the complete transcriptional response of transcription factor deletion. Of the set of KO-regulated targets for a single transcription factor, we wished to detect how many were the result of the primary genetic perturbation and what proportion were secondary effects that were transitively passed from deletion through primary transcription factor targets through a regulatory cascade. Additionally, we wished to characterize both the prevalence and depth of regulatory cascades within the network $G(N)$.

Regulatory Cascade Detection

Regulatory cascades were detectable through topological analysis of the directed and weighted properties of the network. The strategy that we employed is described by Figures 4.6 and 4.7. Figure 4.6 demonstrates the global set of operations used to produce the final regulatory network $G(RN)$. The individual sub-networks $G(KO^n)$ were defined for each transcription factor and its set of KO-regulated targets. These networks were integrated into the combined network $G(N)$. The network $G(N)$ was then evaluated for the presence of regulatory cascades. This operation was performed to both separate secondary regulatory effects from the *G*(*RN*) network and characterize the prevalence and depth of the regulatory cascades. Figure 4.7 clarifies the actual mechanism of candidate cascade evaluation and elimination through descriptive example.

A pair of vertices connected by a directed edge that share a directed third target vertex represent the topological definition of a detected regulatory cascade. The first two vertices are by definition transcription factors, as they are both known to target a third shared vertex. Two criteria were evaluated to determine if the shared target vertex was the result of a regulatory cascade from the first vertex's regulatory influence on the second. First, the directionality of regulation upon the shared vertex must be consistent. Each must either activate or repress the shared target. This was evaluated through modeling the regulatory network *G*(*N*) using the signed (X) score set of values. Second, the significance of the putative secondary regulatory interaction emanating from the first to the shared third vertex had to be measured as less than the significance of the interaction of the second to third vertex. This criteria was evaluated by comparison of *P values* using a *G*(*N*) network constructed with weighted *P value* edges.

Regulatory Cascade Results

The initial unrefined network contained 14,427 total interactions between all transcription factors profiled and significant KO-regulated targets. Regulatory refinement allowing one level of indirect regulation reduced the count to 14,274 interactions. A second and third iteration of the refinement algorithm allowing two and three levels of indirect regulation reduced the interaction counts to 14,258 and 14,251, respectively. Further depth of the refinement procedure was unable to detect qualified secondary regulatory relationships for elimination. As visualized in Figure 4.8, the final refined network *G*(*RN*) was used for all subsequent data analysis.

Figure 4.6 – Network Refinement Process

(a) The process of network refinement begins with modeling the KO-regulated targets of individual transcription factor deletions as separate directed weighted graphs. Arrows represent activation while T structures represent repression. (b) The network $G(N)$ was defined as the union of all nodes and regulatory edges of the individual sub-networks into a single unrefined network. (c) The initial iteration of network refinement demonstrates that *TF A* activates *TF B*. They both share the activated target *Gene M*. If the *P Value* significance of *TF A* activating *Gene M* is lower than the significance *TF B* activating *Gene M* the regulatory interaction between *TF A* and *Gene M* will be designated as secondary and eliminated from the network. (d) Additional refinement steps of two and three levels of cascading regulation were evaluated under the same criteria.

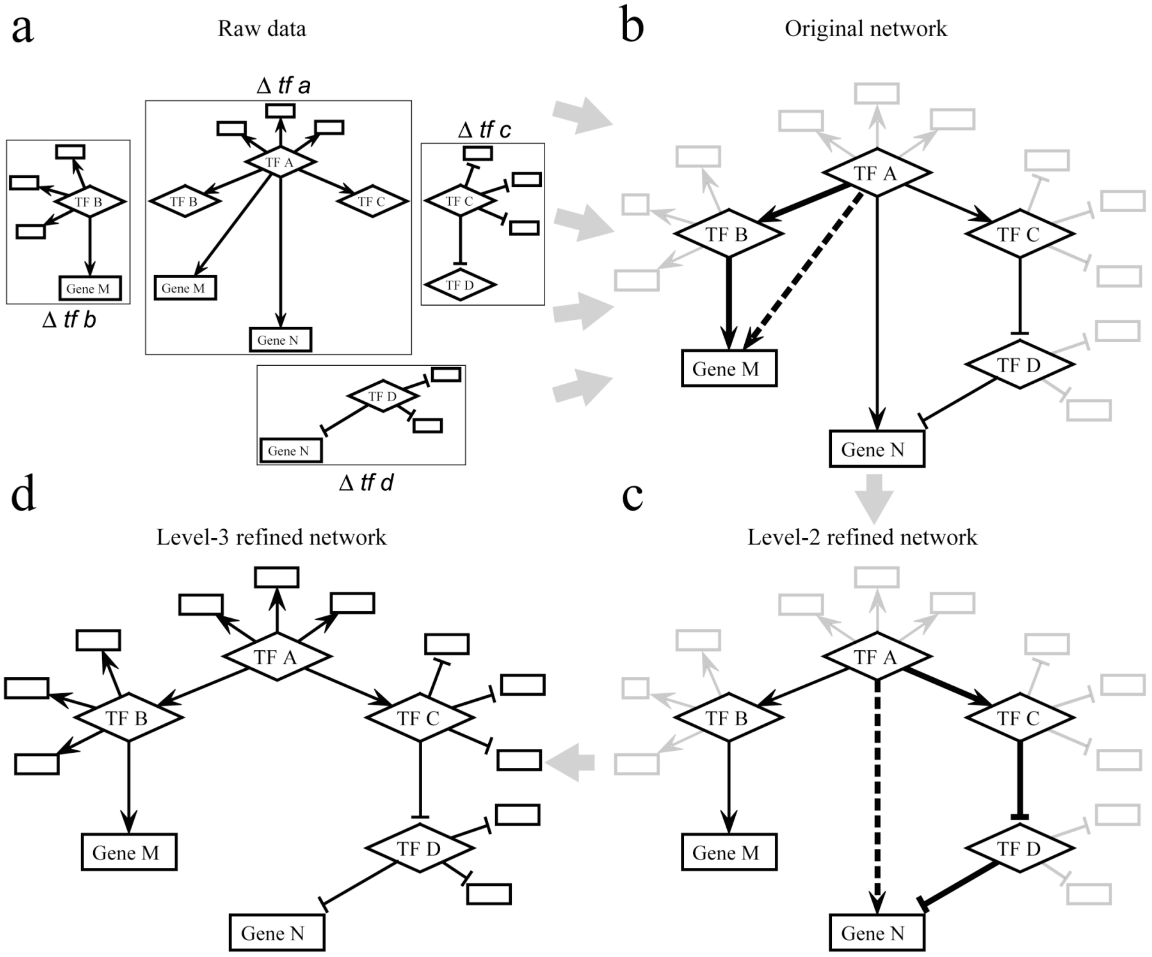
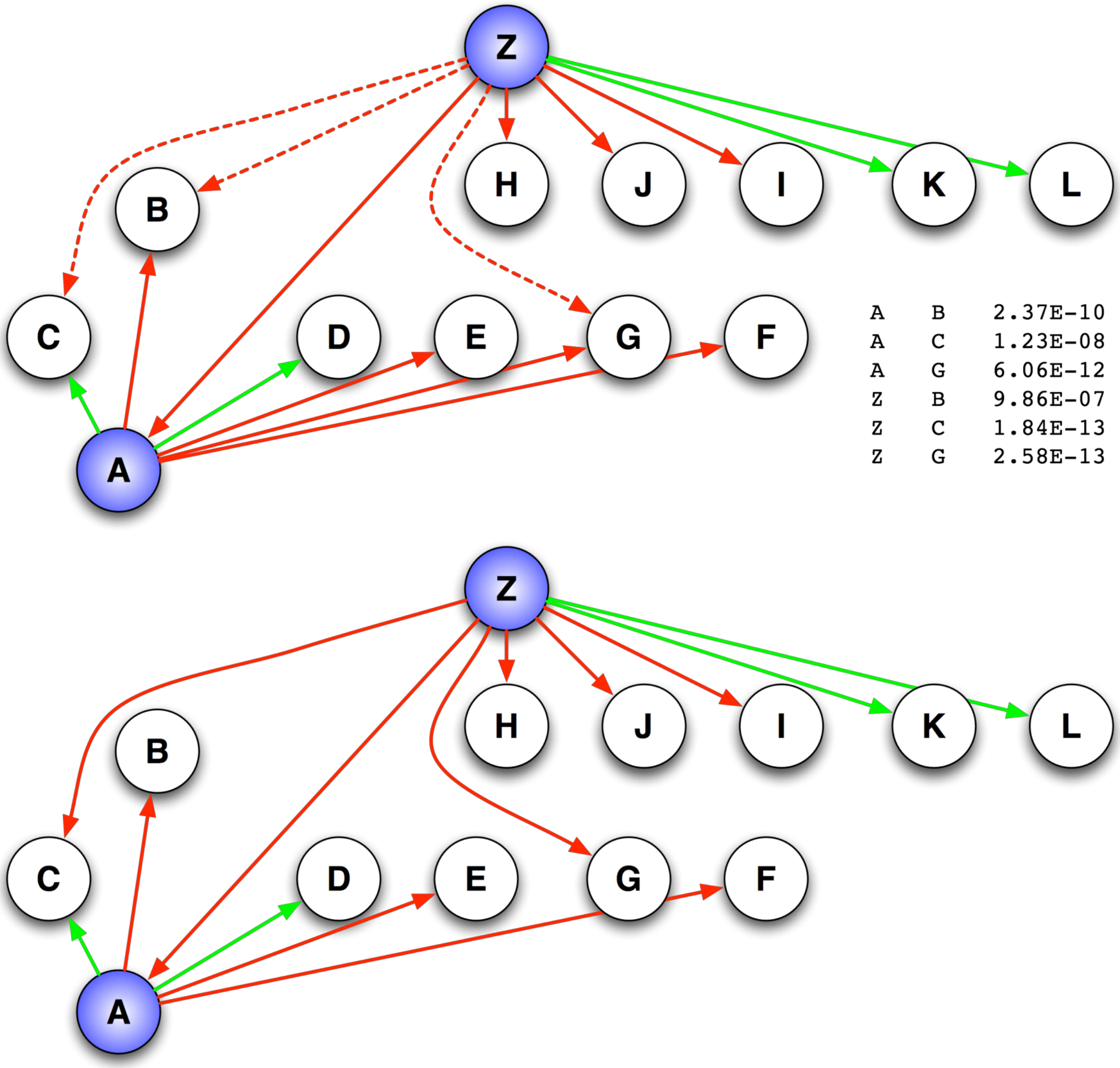


Figure 4.7 – Network Refinement Example

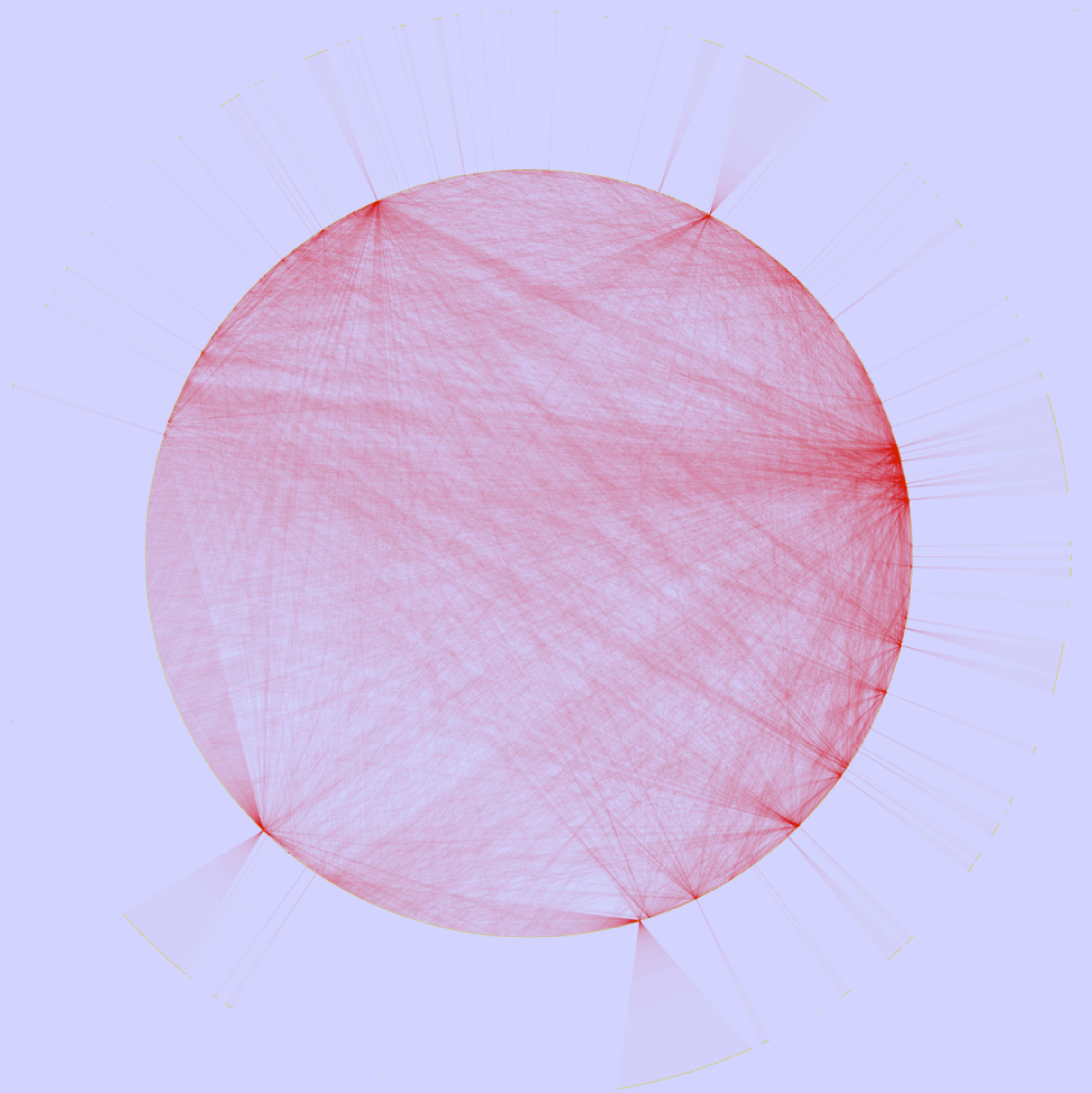
Transcription factor Z is shown to activate transcription factor A. The regulation of Z upon A creates the possibility that each regulatory target that A and Z have in common is primarily regulated by A with Z's regulatory impact being secondary. This possibility is tested for each shared target (dashed lines). The relationship Z◊C is not a candidate for elimination because the directionality of regulation is not consistent between Z◊C and A◊C. The relationships Z◊B and Z◊G have compatible shared directionality of activation as compared to A◊B and A◊G. The *P Value* of Z◊G is more significant than that of A◊G and thus Z◊G is not eliminated. The same evaluation of Z◊B and A◊B shows that Z◊B is a secondary effect and is eliminated.



A	B	2.37E-10
A	C	1.23E-08
A	G	6.06E-12
Z	B	9.86E-07
Z	C	1.84E-13
Z	G	2.58E-13

Figure 4.8 – The Regulatory Network *G*(*RN*)

A visualization of the functional regulatory network *G*(*RN*) was produced by converting the interaction file format to a network-ready format compatible with Cytoscape [51]. This network, specifically enriched for primary targets of transcription factors, was used for all subsections of data analysis.



Ontological Enrichment Analysis

The Gene Ontology initiative is a collaborative effort to provide standardized nomenclature and universal identifiers to describe gene products. The project spans many model organisms including *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Mus musculus*, and *Homo sapiens*. The primary goal of the consortium is the development and maintenance of a structured set of ontologies that describe gene products in terms of their constituent biological processes, cellular components, and molecular functions. GO annotations are curated for the yeast genome by the *Saccharomyces Genome Database* (SGD). We utilized these resources in an effort to further understand the biological theme or set of themes represented by the KO-regulated targets for each transcription factor deletion.

The primary goal of our GO enrichment analysis was to determine whether the deletion of a transcription factor produced a set of KO-regulated targets that have a common biological theme with regards to the analysis of their aggregated GO annotations. This question was addressed in terms of ontologically assigned cellular locations, biological processes, and molecular functions. We analyzed KO-regulated target sets of GO terms statistically for overrepresentation and topologically with regards to their relative placement within the overall GO network. The standard and slim versions of the GO ontologies were used throughout the analysis process.

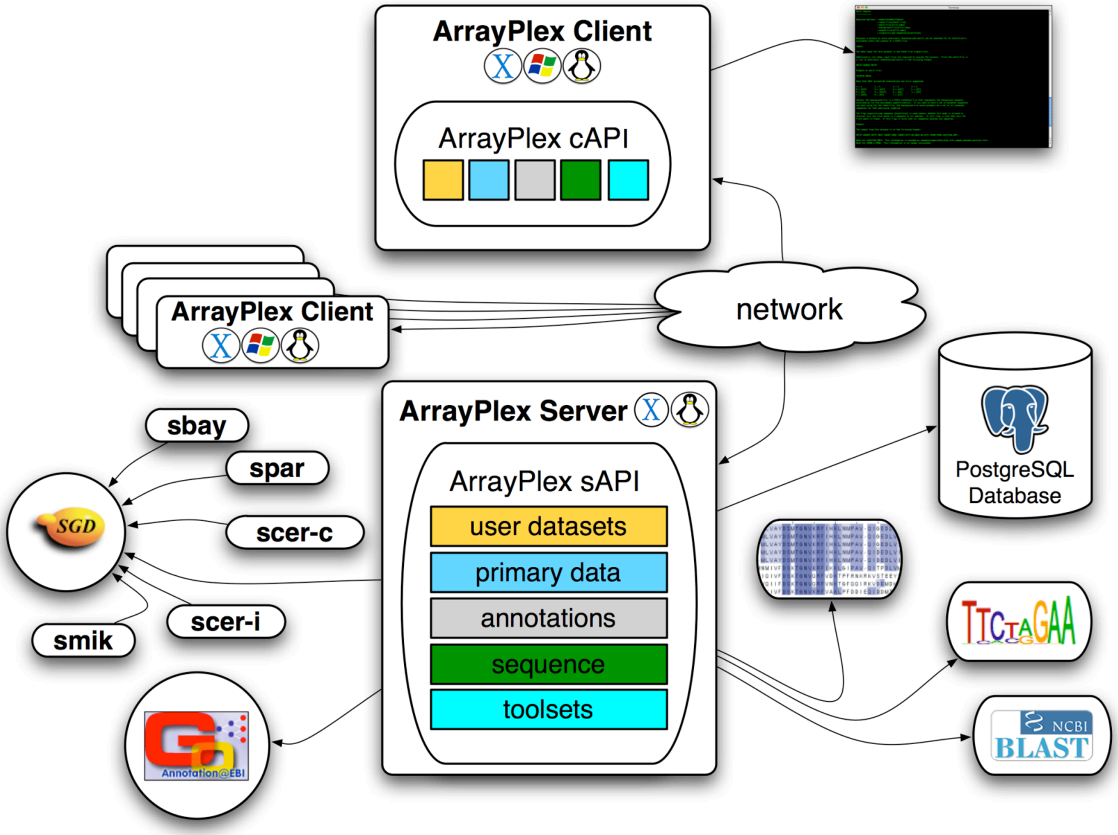
Gene-Specific Assignment Acquisition

The GO enrichment analysis process began with the acquisition of organism-specific GO term assignments for all genes in the yeast genome. ArrayPlex, depicted in Figure 4.9 and described in Chapter 3, provided this information.

Figure 4.9 – Ontology & Genome Sequence Retrieval

The ArrayPlex server environment actively maintains both genomic sequence information and SGD-curated GO term annotations. ArrayPlex Command-Line Modules used the Client API to retrieve this information for both the GO term enrichment and promoter sequence analysis components of KO-regulated target analysis.

ArrayPlex also contains ORF and intergenic alignments for several *sensu stricto Saccharomyces* species. These resources were used during promoter sequence analysis for the purpose of phylogenetic shadowing of candidate cis-regulatory sequences.



Ontological Network Modeling

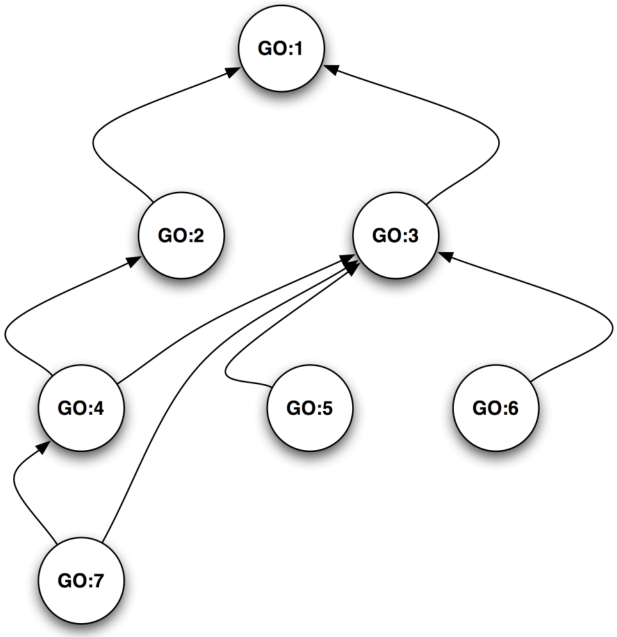
We used directed weighted graphs to model the GO ontologies into networks in much the same way we defined transcription factor sub-networks $G(KO^n)$. The GO ontology networks were built using relational constructs native to the GO nomenclature. GO is explicitly hierarchical; low-level nodes in the set of annotations specify very specific cellular components, biological processes, or molecular functions. These nodes connect upwards through *is_a* or *part_of* relationships to less-specific, more generalized levels of annotation. These sets of relationships were modeled as a directed weighted graph as demonstrated in Figure 4.10. The complete GO ontology is represented in directed graph form as:

$V(GO)$ = { All GO annotations curated for the *Saccharomyces cerevisiae* genome }
This information was provided by SGD.
$E(GO)$ = { All *is_a*, *part_of* relationships }
This information is species-independent and provided by GO Consortium.

Each gene product annotated by SGD was likely to have more than one annotation associated with it. Often a gene will have several annotations in each of the three GO sub-hierarchies (component, process, function). Figure 4.11 demonstrates that each of the KO-regulated targets of a transcription factor has often not just one but a set of assigned GO annotations. Additionally, Figure 4.11 shows that annotation assignments are not unique to single genes. This fact is the basis by which we can search for enrichment of specific GO terms within the KO-regulated targets of individual transcription factor deletions.

Figure 4.10 – GO Ontology Network Modeling

The GO ontology was modeled using directed weighted graphs. The theoretical GO sub-network demonstrates the hierarchical nature of the GO sub-networks. Low-level annotations describing specific biological components, processes, or functions link to higher, more generic annotations. These linkages are directional. As the example demonstrates, lower-level annotations can connect up to multiple levels in the GO sub-networks.



The specific example shows a small subsection of the GO component tree demonstrating specific parts of the proteasome separately annotated and hierarchically related.

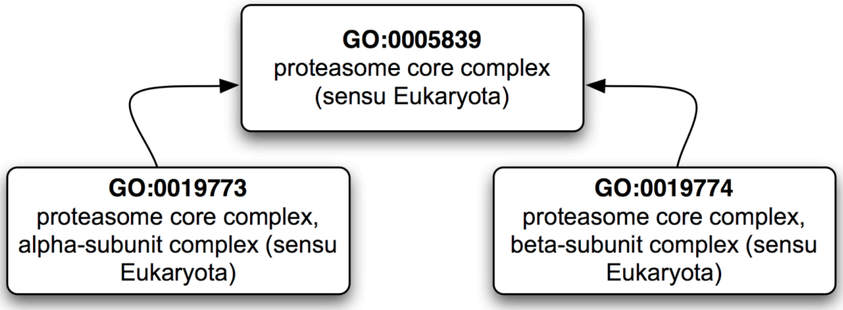
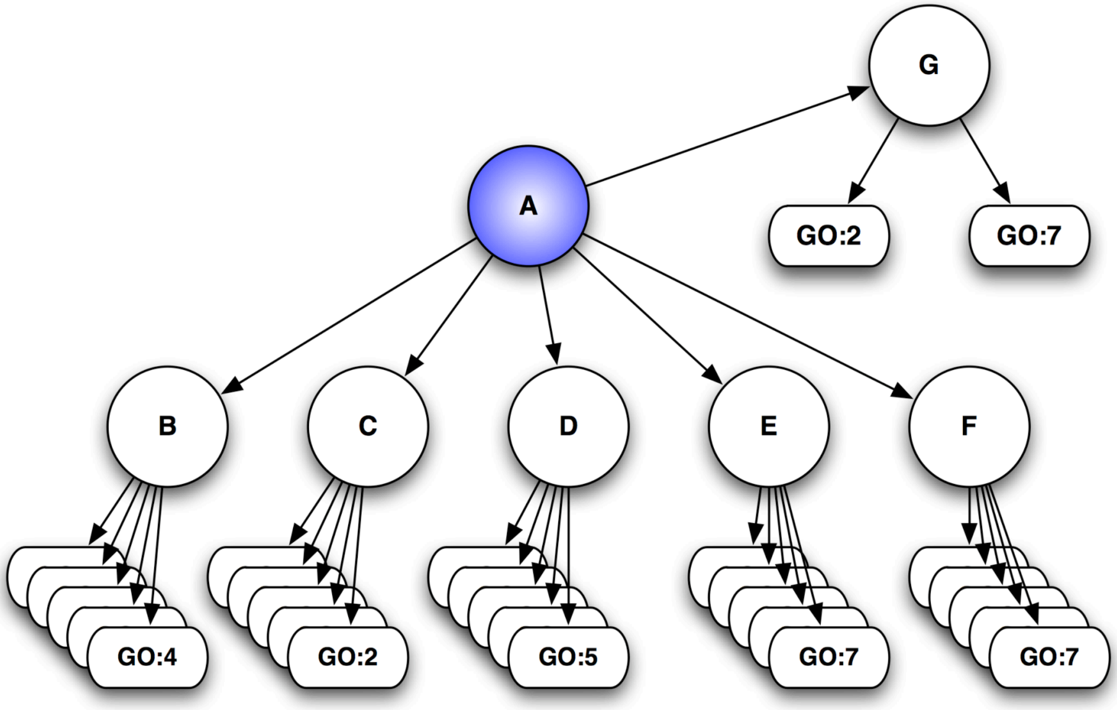

GO:0005839
proteasome core complex (sensu Eukaryota)
GO:0019773
proteasome core complex, alpha-subunit complex (sensu Eukaryota)
GO:0019774
proteasome core complex, beta-subunit complex (sensu Eukaryota)

Figure 4.11 – Transcription Factor Ontology Network

Each KO-regulated target (B, C, D, E, F, G) of transcription factor A has a set of associated GO annotations. These GO annotations can be unique to an individual gene target; many are likely shared among multiple gene targets. In this example, KO-regulated targets E, F, and G share theoretical GO annotation "GO:7".



Raw Term Enrichment

Each transcription factor-specific sub-network $G(KO^n)$ was analyzed for the overrepresentation of individual GO terms. This analysis was performed utilizing the cumulative hypergeometric probability distribution, a discrete probability distribution that describes the number of successes in a sequence of draws from a finite population without replacement.

All calculations were performed according to the unique occurrence of GO annotations for a given SGD-annotated gene product. Each gene product often has an individual GO term associated with it multiple times in the GO ontologies maintained by SGD. The catalog of GO term assignments is maintained with a descriptor termed the evidence code. A single gene product may have multiple entries for the same GO annotation if there are multiple lines of experimental or computational evidence supporting the annotation assignment. For all hypergeometric calculations we were careful to count only unique associations of a GO identifier (GOID) with a SGD identifier (SGDID).

The population size (N) was defined as the total number of unique GOID and SGDID combinations within the entire pool of annotations provided by SGD. The sample size (n) was defined as the total number of unique GOID and SGDID combinations in the pool of annotations for a single transcription-factor specific sub-network $G(KO^n)$. The parameter D, often denoted as the number of successes, is the total number of times a GOID is uniquely associated with a SGDID within the population N. Finally, the parameter k (the match) is the total number of times a GOID is associated with a unique SGDID within the transcription-factor specific sub-network $G(KO^n)$ set of annotations. With each of these parameters established, the probability of GO term overrepresentation was calculated using the cumulative hypergeometric probability function.

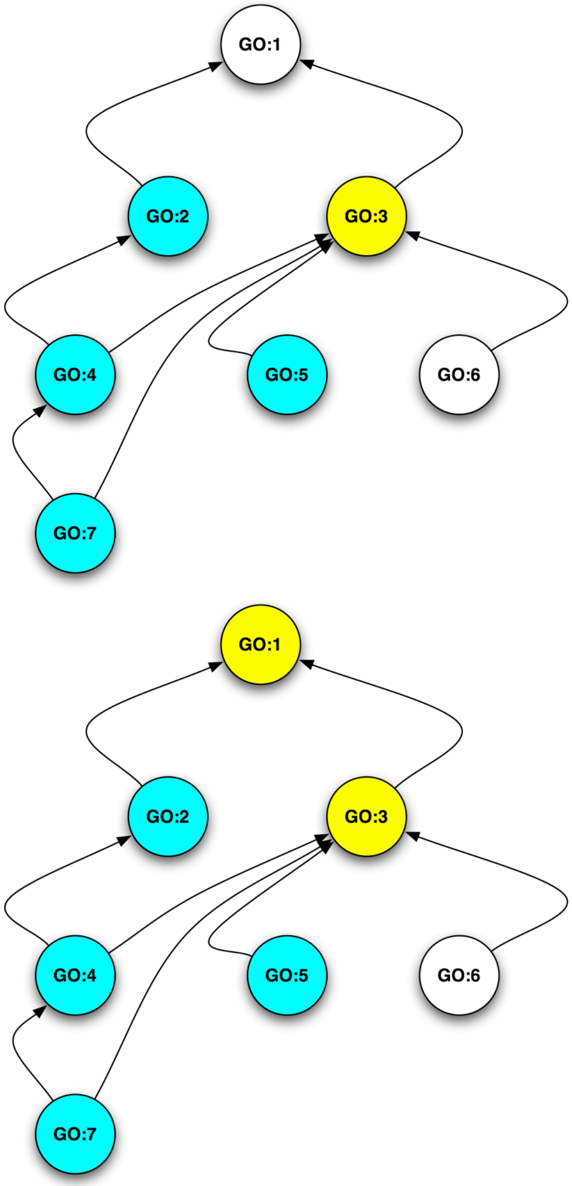
Composite Term Enrichment

The single-term GO analysis procedure uncovered many notable overrepresented annotations for the $G(KO^n)$ sub-networks. We theorized that while a single GO annotation may barely miss the threshold of probabilistic significance, the aggregation of raw annotations upwards within the GO hierarchy may uncover what might otherwise be overlooked biological insights. Figure 4.12 demonstrates this concept.

The annotations GO:2, GO:4, GO:5, and GO:7 are within the pool of annotations provided by some $G(KO^n)$ sub-network. The annotations GO:1, GO:3, and GO:6 are not represented by any of the KO-regulated targets within $G(KO^n)$. While GO annotations GO:2, GO:4, GO:5, and GO:7 are not statistically significant these annotations hierarchically aggregate to the candidate GO term GO:3. Evaluation of the statistical significance at this level GO:3 requires a description of how the hypergeometric parameters must be calculated for this type of analysis. The population size (N) was calculated as previously discussed. The sample size (n), the number of successes (D), and the match (k) were all calculated by summating the values, as they would be calculated for raw analysis of GO annotations GO:1, GO:3, and GO:6. Additionally, n and D values are aggregated for any lower-level GO node (GO:6 in this instance) that is not part of the pool dictated by $G(KO^n)$. The k value for GO node GO:6 is set to zero and does not contribute to additional match. The composite GO annotation GO:3 is created and evaluated for statistical significance. The composite GO annotation GO:1 is created in the secondary round of iteration as an aggregation of GO nodes GO:2 and GO:3.

Figure 4.12 – Composite GO Analysis

Raw Annotations (cyan) are GO nodes that are annotations provided by KO-regulated targets in a sub-network $G(KO^n)$. *Composite Annotations* (yellow) are those uncovered through iterative aggregation of combinations of *raw* annotations and *composite* annotations.



Enrichment Results

The cumulative hypergeometric probability was evaluated for all evaluated GO terms at a stringent Bonferroni corrected statistical cutoff of 4.0 x 10^{-05}. At this threshold 156 transcription factors had significantly enriched GO terms. Of 1113 total transcription factor and GO term combinations, 213 total unique terms were enriched. Of these unique terms, acid phosphatase activity, amino acid catabolism and metabolism, ribosomal biogenesis and constituent components, stress responses, and other metabolic activities were common.

A total of 418 transcription factor and GO Slim term combinations were enriched. Within this pool of combinations, 115 unique transcription factors and 48 GO Slim terms were unique. These GO Slim terms clearly showed that amino acid metabolism, ribosome-centric processes and components, and stress responses were common biological themes that described the KO-regulated set of targets for transcription factor deletions.

The GO term enrichment results are available within the supplementary information section of the publication authored for this study [46]. The GO slim enrichment results are provided in Appendix II.

Term Enrichment Validation

Comparison of enriched GO and GO Slim terms to the SGD annotations for well-characterized transcription factors demonstrated that the analytical procedure recovered known biological components, processes, and functions in which factors were previously shown to function. Transcription factors RPN4, TEC1, RAP1, HSF1, and GCR1 are but a few of the factors whose GO term enrichments correlated well with their known regulatory roles.

Many previously uncharacterized transcription factors showed enrichment for GO terms. Table 4.1 provides the GO terms that were enriched for the transcription factors AFT1 and RTG3. These factors had a relatively small amount of annotation and known regulatory function when queried in SDG. The deletion of AFT1 resulted in enrichment for several ribosomal, stress response, and chaperone terms. Similarly, the deletion of RTG3 showed enrichment for amino acid biosynthesis and transport functions. Specifically, RTG3 was shown to be an activator of glutamate biosynthesis. We wished to validate both the analytical procedure of GO term enrichment and the candidate novel characterizations provided by these enrichments through experimental growth assay.

The transcription factors selected for this analysis were AFT1, RTG3, BAS1, RIC1, and PHO2. The factors BAS1 and PHO2 have previously characterized roles in histidine biosynthesis. RIC1 plays a role in rRNA transcription and promotes the synthesis of other ribosomal proteins. The results of this analysis are demonstrated in Figure 4.13.

The deletion strain *aft1Δ* was plated on rich YPD medium by serial dilution and was shown to exhibit growth similar to wildtype at 30°C. The *aft1Δ* strain, however, exhibited a pronounced growth defect at the heat-stress condition 37°C. The deletion strains *rtg3Δ*, *bas1Δ*, *ric1Δ*, and *pho2Δ* were serially diluted and plated on both rich YPD medium and synthetic minimal medium with dextrose, uracil, histidine, methionine, and leucine. Of these strains, only *rtg3Δ* showed a significant growth defect on minimal medium lacking glutamate.

Table 4.1 – GO Term Enrichment For Growth Assays

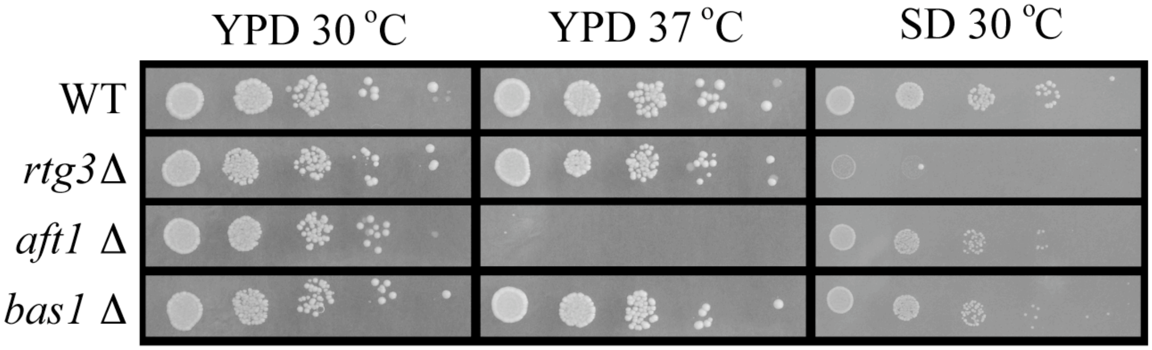
TF Systematic	TF Symbol	GO ID	*P Value*	Term
YBL103C	RTG3	GO:0000943	2.52E-37	retrotransposon nucleocapsid
YBL103C	RTG3	GO:0003723	1.76E-26	RNA binding
YBL103C	RTG3	GO:0005515	2.68E-20	protein binding
YBL103C	RTG3	GO:0003676	1.07E-12	nucleic acid binding
YBL103C	RTG3	GO:0016813	3.86E-06	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amidines
YBL103C	RTG3	GO:0019878	4.01E-06	lysine biosynthesis via aminoadipic acid
YBL103C	RTG3	GO:0005275	4.74E-05	amine transporter activity
YBL103C	RTG3	GO:0000256	7.44E-05	allantoin catabolism
YBL103C	RTG3	GO:0005488	1.13E-04	binding
YBL103C	RTG3	GO:0006536	1.55E-04	glutamate metabolism
YBL103C	RTG3	GO:0004410	2.46E-04	homocitrate synthase activity
YBL103C	RTG3	GO:0005371	2.46E-04	tricarboxylate carrier activity
YBL103C	RTG3	GO:0003994	2.46E-04	aconitate hydratase activity
YBL103C	RTG3	GO:0046912	4.26E-04	transferase activity, transferring acyl groups, acyl groups converted into alkyl on transfer
YBL103C	RTG3	GO:0015291	6.15E-04	porter activity
YBL103C	RTG3	GO:0016836	7.93E-04	hydro-lyase activity
YBL103C	RTG3	GO:0015171	8.35E-04	amino acid transporter activity
YBL103C	RTG3	GO:0006537	9.81E-04	glutamate biosynthesis
YGL071W	AFT1	GO:0005830	4.46E-28	cytosolic ribosome (sensu Eukaryota)
YGL071W	AFT1	GO:0005842	5.01E-22	cytosolic large ribosomal subunit (sensu Eukaryota)
YGL071W	AFT1	GO:0043037	8.77E-22	translation
YGL071W	AFT1	GO:0003735	1.27E-19	structural constituent of ribosome
YGL071W	AFT1	GO:0005198	2.24E-16	structural molecule activity
YGL071W	AFT1	GO:0005843	1.14E-08	cytosolic small ribosomal subunit (sensu Eukaryota)
YGL071W	AFT1	GO:0006457	4.64E-08	protein folding
YGL071W	AFT1	GO:0009987	1.08E-07	cellular process
YGL071W	AFT1	GO:0005829	1.21E-07	cytosol
YGL071W	AFT1	GO:0006650	1.53E-07	glycerophospholipid metabolism
YGL071W	AFT1	GO:0006616	4.61E-07	SRP-dependent cotranslational protein-membrane targeting, translocation
YGL071W	AFT1	GO:0009066	7.84E-07	aspartate family amino acid metabolism
YGL071W	AFT1	GO:0000788	2.71E-06	nuclear nucleosome
YGL071W	AFT1	GO:0016282	7.20E-06	eukaryotic 43S preinitiation complex
YGL071W	AFT1	GO:0051082	1.40E-05	unfolded protein binding
YGL071W	AFT1	GO:0050875	1.40E-05	cellular physiological process
YGL071W	AFT1	GO:0006096	3.11E-05	glycolysis
YGL071W	AFT1	GO:0000027	3.13E-05	ribosomal large subunit assembly and maintenance
YGL071W	AFT1	GO:0046688	8.15E-05	response to copper ion
YGL071W	AFT1	GO:0009277	9.67E-05	cell wall (sensu Fungi)
YGL071W	AFT1	GO:0006950	1.09E-04	response to stress
YGL071W	AFT1	GO:0006006	1.39E-04	glucose metabolism
YGL071W	AFT1	GO:0006333	1.68E-04	chromatin assembly or disassembly
YGL071W	AFT1	GO:0005788	2.32E-04	endoplasmic reticulum lumen
YGL071W	AFT1	GO:0006644	2.32E-04	phospholipid metabolism
YGL071W	AFT1	GO:0030003	5.00E-04	cation homeostasis
YGL071W	AFT1	GO:0009092	5.14E-04	homoserine metabolism
YGL071W	AFT1	GO:0007582	5.54E-04	physiological process
YGL071W	AFT1	GO:0006979	6.93E-04	response to oxidative stress
YGL071W	AFT1	GO:0005199	9.29E-04	structural constituent of cell wall

Figure 4.13 – Condition-Specific GO Term Enrichment Validation

The deletion strains *rtg3Δ*, *aft1Δ*, *bas1Δ*, *ric1Δ,* and *pho2Δ* were serially diluted and plated on rich YPD medium and synthetic minimal medium with dextrose, uracil, histidine, methionine, and leucine.

The transcription factor AFT1 was predicted to regulate the expression of chaperone proteins. Growth of the deletion strain aft1Δ at heat-stress conditions of 37°C produced the expected growth defect.

The transcription factors RTG3, BAS1, RIC1, and PHO2 were similarly grown on both YPD and minimal medium (dextrose, uracil, histidine, methionine, and leucine). RTG3 was predicted to play a role in the promotion of glutamate biosynthesis. BAS1 and PHO2 were previously characterized to play roles in histidine biosynthesis while RIC1 has been shown to play a role in rRNA transcription and ribosomal protein genes. Of these factors only RTG3 showed a growth defect under minimal medium lacking glutamate.



Network Analysis of Enriched Terms

We utilized graph traversal and network-descriptive measures to quantify the results of our ontological analysis. The input to the process was the results of the *Raw* and *Composite* GO term enrichment analysis. The GO hierarchy is composed of three independent and disconnected hierarchies: the biological process tree, the molecular function tree, and the cellular component tree. Each of the analysis methodologies detailed was performed within the context of an enriched GOID's designated GO tree.

GO sub-hierarchies were modeled using undirected non-weighted graphs. The use of an undirected network allowed us to explore the relative distances and relationships between GOIDs while intentionally disregarding the top-down and directionally hierarchical nature of the sub-hierarchies. Formal definitions of the graph model constituents are documented in Table 4.2.

Table 4.2 – Graph Definitions for GO Network Analysis

Formal definitions of graph model constituent elements used in enriched GO term network analysis traversal and analysis.

$\mathrm{V}(GO_X)$	{ All GO annotations with aspect $X=P \mid F \mid C$ }
$\mathrm{E}(GO_X)$	{ All *is_a*, *part_of* relationships for the GOID nodes in $\mathrm{V}(GO_X)$ }
$\mathrm{V}(TF_{X,Y})^{\mathrm{RAW}}$	{ All RAW GO annotations with aspect $X=P \mid F \mid C$ statistically significant for TF=Y }
$\mathrm{E}(TF_{X,Y})^{\mathrm{RAW}}$	{ All *is_a*, *part_of* relationships for the GOID nodes in $\mathrm{V}(TF_{X,Y})^{\mathrm{RAW}}$ }
$\mathrm{V}(TF_{X,Y})^{\mathrm{RAW+COMPOSITE}}$	{ All RAW+COMPOSITE GO annotations with aspect $X=P \mid F \mid C$ statistically significant for TF=Y }
$\mathrm{E}(TF_{X,Y})^{\mathrm{RAW+COMPOSITE}}$	{ All *is_a*, *part_of* relationships for the GOID nodes in $\mathrm{V}(TF_{X,Y})$ }
$\mathrm{V}(TF_{X,Y})^{\mathrm{RAW,RANDOM}}$	{ Random set where member size matches $\mathrm{V}(TF_{X,Y})^{\mathrm{RAW}}$ }
$\mathrm{E}(TF_{X,Y})^{\mathrm{RAW,RANDOM}}$	{ All *is_a*, *part_of* relationships for the GOID nodes in $\mathrm{V}(TF_{X,Y})^{\mathrm{RAW,RANDOM}}$ }
$\mathrm{V}(TF_{X,Y})^{\mathrm{RAW+COMPOSITE,\ RANDOM}}$	{ Random set where member size matches $\mathrm{V}(TF_{X,Y})^{\mathrm{RAW+COMPOSITE}}$ }
$\mathrm{E}(TF_{X,Y})^{\mathrm{RAW+COMPOSITE,\ RANDOM}}$	{ All *is_a*, *part_of* relationships for the GOID nodes in $\mathrm{V}(TF_{X,Y})^{\mathrm{RAW+COMPOSITE,\ RANDOM}}$ }

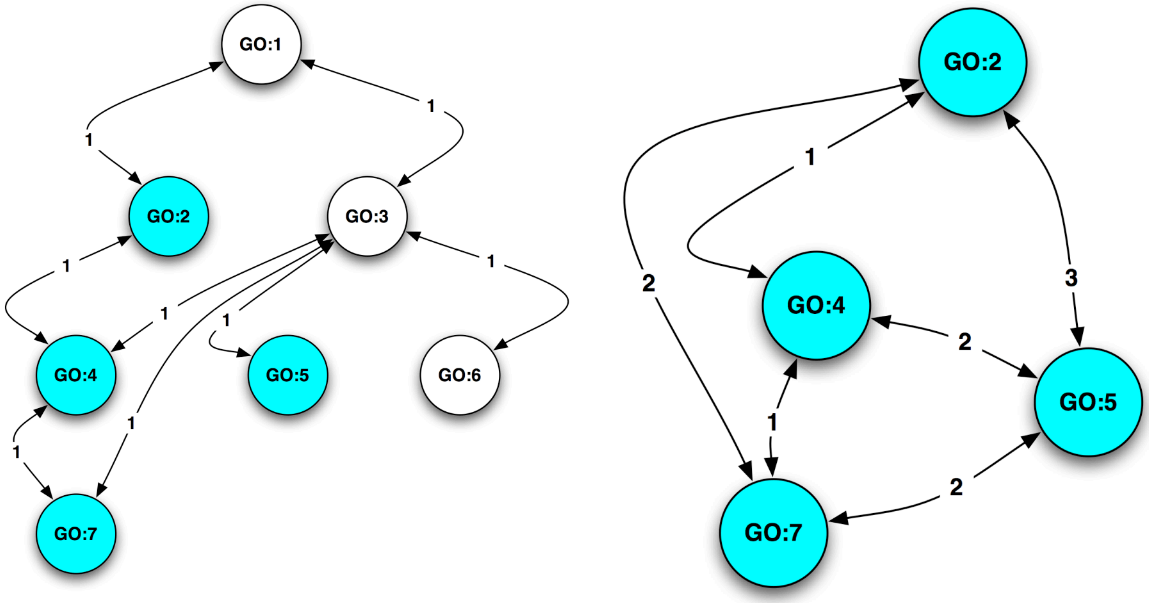
Network Metric, Average Path Length

The first quantification methodology focused upon the analysis of the average path length between enriched GO annotations. For each transcription factor deletion profiled, the statistically significant *Raw* GOID vertices $\mathrm{V}(TF_{X,Y})^{\mathrm{RAW}}$ were tagged within their modeled GO sub-hierarchy $\mathrm{V}(GO_X)$. All tagged vertices were then analyzed in a pair-wise fashion in order to determine the complete set of all possible minimum shortest path distances between GOID vertex combinations traversing edge elements in $\mathrm{E}(TF_{X,Y})^{\mathrm{RAW}}$. The arithmetic mean of these values was calculated to determine an *average path length* between the set of GOIDs. This process is detailed in Figure 4.14.

This process was then iteratively performed for random permutations of node enrichments (using $\mathrm{V}(TF_{X,Y})^{\mathrm{RAW,RANDOM}}$, $\mathrm{E}(TF_{X,Y})^{\mathrm{RAW,RANDOM}}$). For each set of GOIDs, equivalently sized sets of random SGD-curated GOIDs were tagged to appropriate GO sub-hierarchies. Graph traversal algorithms were similarly used to determine pair-wise distances sufficient for the calculation of random average path lengths. Iterations sets of both 20 and 100 in size were used to directly measure the number of times the actual average path length was less than the random path length. The set of 100 random iterations provided a statistical foundation by which *P Values* could be derived describing the probabilistic expectation that the experimental average path length is significantly and reliably less than the random average path length. Finally, each of the procedures detailed above was performed for the combination of RAW+COMPOSITE GOID annotations determined to be significant for each transcription factor KO (using $\mathrm{V}(TF_{X,Y})^{\mathrm{RAW+COMPOSITE}}$, $\mathrm{E}(TF_{X,Y})^{\mathrm{RAW+COMPOSITE}}$, $\mathrm{V}(TF_{X,Y})^{\mathrm{RAW+COMPOSITE,RANDOM}}$, $\mathrm{E}(TF_{X,Y})^{\mathrm{RAW+COMPOSITE,RANDOM}}$).

Figure 4.14 – Average Path Length

Determination of average path length for a set of statistically significant GOIDs. All shortest-path distances for all possible pair-wise combinations of enriched nodes (cyan) were determined using conventional graph traversal algorithms. The arithmetic mean of these pair-wise distances provides a means by which the average network distance between enriched nodes can be determined. The average path length for the above example would be 1.83.



Network Metric, Clustering Coefficient

The clustering coefficient was first developed to demonstrate whether or not a graph might be considered to be a small-world network. Formally defined, the clustering coefficient for a chosen vertex of a graph is the proportion of edges between the selected vertex and the vertices to which it is directly connected divided by the total number of edges that could possibly connect them all together.

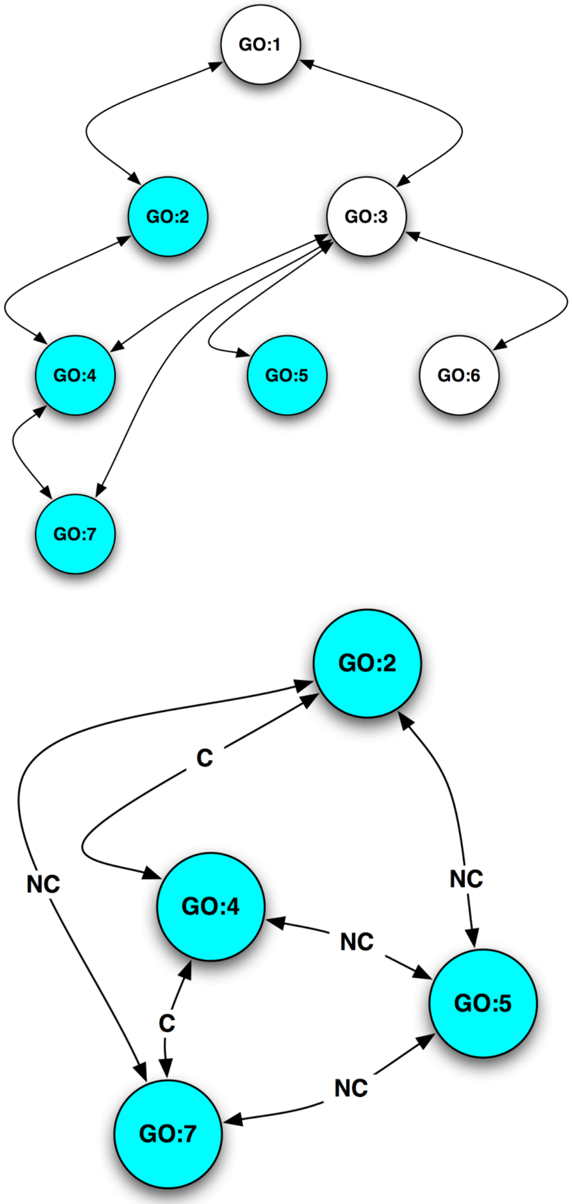
Mathematical definition of clustering coefficient C_i for some vertex V_i in the set of vertices $V(TF_{X,Y})^Z$ where Z can take on values of (RAW; RAW+COMPOSITE; RAW, RANDOM; RAW+COMPOSITE, RANDOM).

$$C_i = \frac{2 * \left| E\left(TF_{X,Y}\right)^Z \right|}{V\left(TF_{X,Y}\right)^Z * \left(V\left(TF_{X,Y}\right)^Z - 1\right)}$$

The procedures executed directly mirror those detailed in the average path length analysis. The clustering coefficient is evaluated for each statistically significant *Raw* GOID of every transcription factor deletion profiled.

Figure 4.15 – Clustering Coefficient with GO Term Enrichment

Determination of clustering coefficient for a set of statistically significant GOIDs. Edge tags of NC and C show the ability to not determine and determine a non-zero clustering coefficient respectively.



Network Analysis Results

For GO terms enriched at *P Value* less than 0.001 more than 70% of the terms demonstrated average path lengths that were statistically smaller than random permutations (*P Value* less than 0.01). Similarly, more than 65% of the transcription factors with two or more enriched GO terms in the same GO sub-network (component, process, function) have significant clustering coefficients as compared with random assignment and evaluation.

These results were supportive of the overall results for GO term enrichment analysis. The smaller than random average path lengths demonstrated co-enrichment of terms from co-localized areas of the GO hierarchies. Similarly, statistically high clustering coefficients took the average path length results one level of confidence further by demonstrating that co-enriched GO terms were not only co-localized to a general areas of but in many instances shared a direct topological connection with each other.

Target Promoter Sequence Motif Analysis

Core to the process of differential gene expression and cell-fate determination are the cis-regulatory elements that define eukaryotic promoters. Eukaryotic promoters generally contain three common elements, the first two defining what is often called the core promoter: the transcription-start site, the TATA box, and upstream sequences including activators, enhancers, repressors, and silencers [9]. We designed a sequence analysis methodology for the purpose of studying the promoter sequences of KO-regulated targets of each profiled transcription factor. We wished to determine whether a set of KO-regulated targets have statistically significant over-represented sequence motifs in their promoter regions. Additionally, we wished to profile the over-representation of characterized motifs within these regulatory regions.

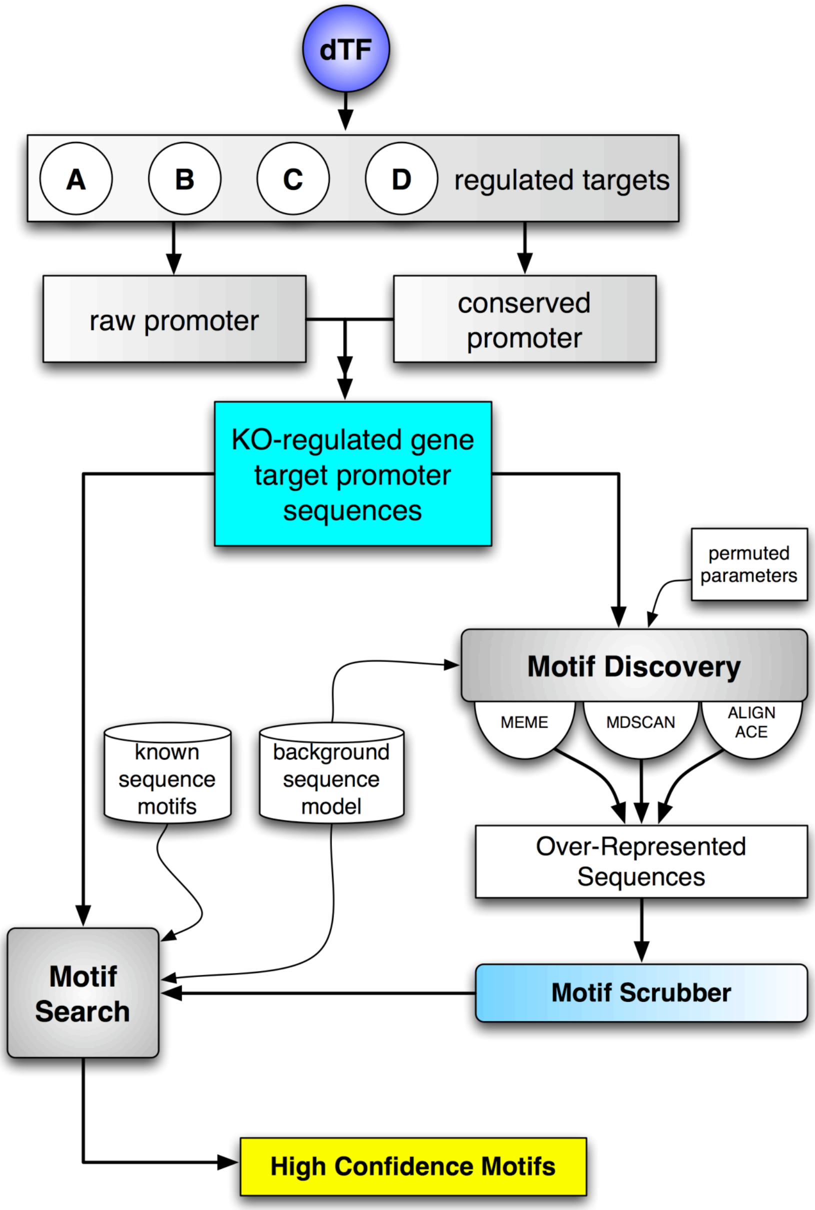
The analysis of promoter sequences required the parametric acquisition of primary sequence data. In order to evaluate the efficacy of phylogenetic shadowing, primary genomic sequence was needed for both *Saccharomyces cerevisiae* as well as other related *sensu stricto Saccharomyces* species. As detailed in Figure 4.9 and discussed in the ontological term enrichment sections of this chapter, this information was provided by the ArrayPlex. ArrayPlex contained primary genomic sequence data for *S. cerevisiae*, *S. mikatae*, and *S. bayanus*. In addition to raw genomic sequence information, ArrayPlex provided access to protein-coding and intergenic alignments between each of these yeast species. These alignments allowed us to analyze promoter sequences utilizing phylogenetic shadowing to select for conserved non-coding regions of cis-regulatory promoter sequence. In addition to primary sequence data, ArrayPlex had the capacity to dynamically execute and return normalized results from three separate software packages (AlignACE, MEME, and MDscan) that have been well characterized in previous efforts to analyze sequence sets for over-represented sequence motifs [REFs].

Sequence Analysis Life Cycle

The sequence analysis process was performed for each KO-regulated target set. Figure 4.16 provides a top-down overview of the process through which each target set passed. A set of statistically significant targets was previously determined for each transcription factor. The relationship between transcription factors and their specific set of targets was expressed in terms of an interaction file, previously described by Figure 4.9. The set of targets for the example transcription factor deletion in Figure 4.16 (dTF) are A, B, C, and D.

Figure 4.16 – Sequence Analysis Life Cycle

Life cycle of the sequence motif analysis process for a single transcription factor KO. Transcription factor deletion (dTF) has a predetermined set of statistically significant targets A, B, C, and D.



Promoter Sequence Extraction

Two sets of sequence information were retrieved for each of the transcription factor KO-regulated targets. First, multi-species *sensu stricto Saccharomyces* 5'intergenic sequence alignments (CLUSTALW format) were retrieved, parsed, and post-processed to produce a set of target-provided promoter sequences in which the sequence information was filtered for phylogenetic conservation. Phylogenetic conservation was guided by user-provided parameters controlling the window-size of sequence evaluation as well as the proportion of nucleotides that must be conserved within a sliding window to tag a specific region as well conserved. The other set of sequence information was the raw *Saccharomyces cerevisiae* promoter sequence from each of the transcription factor KO-regulated targets. Raw sequence extraction was guided by parameters controlling the promoter distance of sequence retrieved. Additionally, the algorithm was tuned to extract only intergenic sequence and not stray into neighboring ORFs. The two separate sets of sequence information were formatted in FASTA format for use in both the search and discovery phases of promoter sequence analysis.

Motif Search

In parallel to the motif discovery, a motif search module was executed to scan the assembled promoter regions for the existence of previously elucidated and characterized cis-regulatory sequences. Again, this process utilized a pre-determined background sequence model to guide the significance of located motifs. The background sequence model used in this process was a pre-computed FASTA file of all *Saccharomyces cerevisiae* promoter regions. A set of consensus sequences for each of the transcription factors deletions was provided. Each of the consensus sequences was used to synthesize a regular expression that was searched for in both the experimental and background FASTA files. The number of promoters with hits found for each consensus sequence was recorded for both the background and experimental sequence sets. The information was assembled and the cumulative hypergeometric probability was calculated to evaluate significance of enrichment.

Motif Search Results

Of the transcription factors profiled, 102 had previously characterized sequence motifs. Table 4.3 shows that 40 of these factors have sequence motifs that were statistically enriched within their KO-regulated target promoter regions.

Table 4.3 – Complete Motif Search Results

The complete results of characterizing the statistical overrepresentation of previously characterized sequence motifs within the promoter regions of KO-regulated targets.

TF Systematic Name	**TF Gene Name**	**Motif**	***P Value***
YBR083W	TEC1	CATTCY	7.21E-03
YDL020C	RPN4	GGTGGCAAA	1.21E-16
YDL056W	MBP1	ACGCGT	4.42E-05
YDL106C	PHO2	TAATRA	3.17E-03
YDR207C	UME6	GCGGC	1.86E-17
YDR310C	SUM1	GYGWCASWAAW	1.71E-16
YDR423C	CAD1	TTACTAA	8.49E-03
YDR451C	YHP1	TAATTG	4.23E-03
YER040W	GLN3	GATAAGATAAG	9.14E-03
YER111C	SWI4	CGCSAAA	4.24E-06
YGL071W	RCS1	GGGTGCANT	3.79E-03
YGL073W	HSF1	GARNNTTCNNGAA	2.66E-14
YGL237C	HAP2	CCAAT	8.60E-04
YHL027W	RIM101	TGCCAAG	5.18E-09
YHR178W	STB5	CGGNSTTATA	2.29E-03
YHR206W	SKN7	ATTTGGCYGGSCC	7.64E-04
YJL056C	ZAP1	ACCYYNAAGGT	6.91E-05
YJL110C	GZF3	GATAAG	1.60E-05
YJR060W	CBF1	CACGTG	2.28E-06
YKL015W	PUT3	CGGNNNNNNNNNNCCG	4.05E-03
YKL043W	PHD1	SCNGCNGG	2.38E-03
YKR099W	BAS1	TGACTC	9.21E-10
YLR014C	PPR1	TTCGGNNNNNNCCGAA	2.86E-03
YLR131C	ACE2	GCTGGT	1.13E-05
YLR176C	RFX1	TCGCCATGGCAAC	2.10E-03
YLR403W	SFP1	AYCCRTACAY	7.39E-46
YLR451W	LEU3	CCGNNNNCGG	1.40E-08
YML007W	YAP1	TTASTMA	8.79E-03
YML051W	GAL80	CGGNNNNNNNNNNNCCG	2.07E-06
YMR019W	STB4	TCGGNNCGA	4.74E-03
YMR021C	MAC1	GAGCAAA	2.15E-03
YMR037C	MSN2	MAGGGG	7.80E-11
YMR070W	MOT3	YAGGYA	1.28E-03
YNL068C	FKH2	GGTAAACAA	7.49E-03
YNL216W	RAP1	CAYCCRTRCA	5.74E-37
YOL028C	YAP7	MTKASTMA	6.32E-03
YOR113W	AZF1	YWTTKCKKTYYCKGYKKY	3.40E-04
YPL049C	DIG1	TGAAACA	2.10E-06
YPL075W	GCR1	CWTCC	3.99E-08
YPR065W	ROX1	YSYATTGTT	9.17E-05

Motif Discovery

Sequence sets for each KO-regulated target pool were submitted to the ArrayPlex server for *de novo* analysis using the packages AlignACE, MEME, and MDscan. Each of these programs is a Gibbs sampler designed to use a pre-determined background sequence model to find over-represented motifs within the set of sequences provided. The background sequence model we utilized was a nucleotide frequency matrix as computed by analysis of all *Saccharomyces cerevisiae* intergenic regions. Each Gibbs sampler was guided by a set of computationally varied parameters including the desired motif width and the number of expected motifs. Output from this process was normalized from the native output of each of the Gibbs samplers to a universal format.

Motif Discovery Results & Aggregation

The normalized information was processed into the relational database for the purpose of high-confidence motif discovery. Motif aggregation was performed by the following procedure. First, the relational database, when populated with all candidate motifs, contained nearly 400,000 motif records. These records were first filtered by a parametrically varied set of tool scores specific to significance thresholds of each individual Gibbs sampler. Next, the records that passed the first set of tool score filters were subjected to nucleotide complexity requirements, motif width limitations, duplicate elimination, as well as negative selection against sub-motifs, reverse-complement motifs, and previously characterized motifs. Finally, the high-confidence candidate motifs that survived all of these filters were subjected to statistical significance quantification and filtered by *P Value* using the motif search methodology previously described.

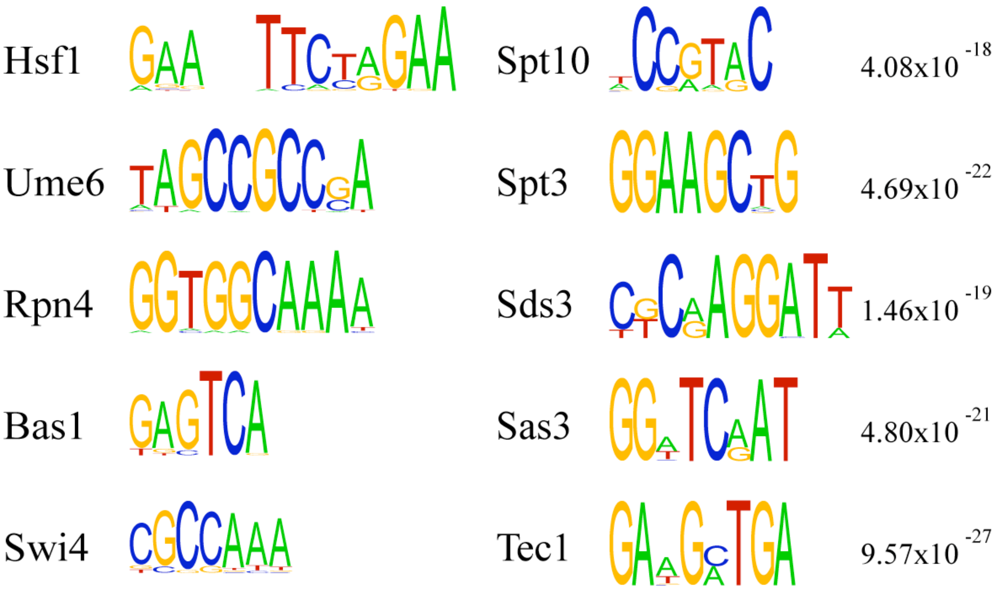
Final Motif Aggregation

The intersection of sequence motifs enriched by the process of directed *motif search* as well as overrepresented by the process of *motif discovery* was aggregated together to form a set of high-confidence *novel* sequence motifs for each KO-regulated target set. A total of 105 transcription factors had 490 unique motifs that passed all of the filters applied in the final aggregation process. A selection of these motifs is presented in Figure 4.17. The complete set of high-confidence novel motifs appear in Appendix III.

Figure 4.17 – Sample Motif Search & Discovery Results

The left column is a selection of previously characterized motifs enriched within the promoter regions of KO-regulated target sets. Table 4.3 contains a complete list of these motifs.

The right column is a selection of novel motifs characterized by the process of motif discovery and final aggregation through a series of strict filters. Appendix III presents a complete list of these motifs.



Binding Does Not Affirm Regulation

For each transcription factor profiled, we first evaluated KO-regulated target sets by comparing them to sets of targets that were determined by previous large-scale studies of yeast transcription factors [47, 48]. These previous studies measured DNA binding events and denoted proximal genes as putative regulatory targets. Our research focused upon the deletion of transcription factors and the determination of targets by identifying transcript levels that were affected by the genetic perturbation. The overlap between our KO-regulated target sets and these previous studies have been previously discussed in this chapter to be low.

It is possible that promoters occupied by transcription factors are not necessarily activating or repressing proximal targets. To investigate this hypothesis further we decided to investigate the effect of transcription factor overexpression, increased binding, and measured RNA transcript levels. We overexpressed the transcription factor HSF1 and determined significant targets. We then analyzed previous studies to determine the set of genes bound by HSF1, induced at least 2.5-fold upon heat-shock, and yet not induced upon HSF1 overexpression.

The result set depicted in Table 4.4 showed us that 28% of the targets occupied by HSF1 and activated by heat-shock were not activated by increasing HSF1 binding. These targets represent a portion of the HSF1 binding targets that require a regulatory step independent of binding to activate proximal targets.

Table 4.4 – HSF1 Post-Binding Regulatory Analysis

The set of genes bound by HSF1, induced at least 2.5 fold upon heat-shock, and not induced upon over-expression of HSF1.

Systematic Name	**Gene Name**	**Heat Shock Induction**	**HSF1 OE Induction**
YBL075C	SSA3	43.41	1.48
YBR053C		17.51	0.96
YBR101C	FES1	10.13	1.27
YCR010C	ADY2	3.07	1.22
YCR011C	ADP1	3.66	0.87
YDR003W	RCR2	2.62	1.34
YDR210W		3.10	1.22
YDR216W	ADR1	3.48	0.73
YDR231C	COX20	2.69	1.06
YDR247W	VHS1	3.16	1.37
YDR258C	HSP78	41.64	0.76
YDR259C	YAP6	3.27	1.11
YER033C	ZRG8	4.72	1.00
YER037W	PHM8	13.36	1.02
YGL036W		4.00	1.09
YGR141W	VPS62	3.61	1.35
YGR250C		12.21	1.18
YHR082C	KSP1	3.73	1.38
YIR017C	MET28	4.26	1.00
YIR038C	GTT1	6.92	1.00
YJL148W	RPA34	2.53	0.79
YJR046W	TAH11	3.86	1.00
YKL010C	UFD4	2.99	1.06
YKL109W	HAP4	8.51	0.73
YKL163W	PIR3	10.93	1.42
YKL164C	PIR1	2.75	1.11
YLL023C		5.58	1.34
YLL039C	UBI4	26.72	1.19
YLR168C		3.84	0.98
YLR218C		2.64	0.98
YLR260W	LCB5	3.34	1.33
YML100W	TSL1	135.30	0.85
YMR251W-A	HOR7	14.52	1.24
YNL007C	SIS1	8.46	0.84
YNL077W	APJ1	20.11	0.99
YNL125C	ESBP6	4.69	1.29
YNL194C		61.39	0.75
YNR069C	BSC5	5.39	1.34
YOR267C	HRK1	4.26	1.38
YPL054W	LEE1	4.47	1.22
YPL250C	ICY2	15.45	1.37
YPR158W		9.19	1.23

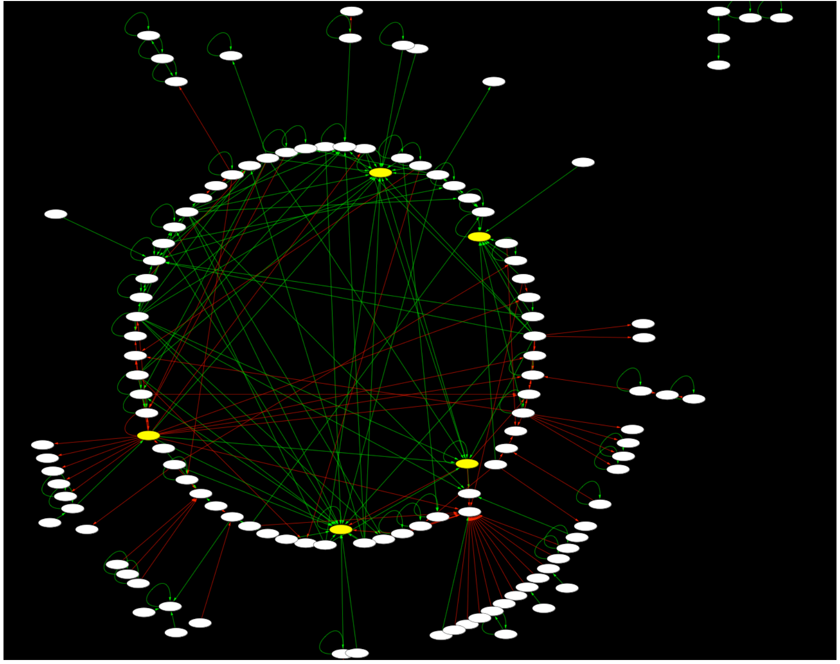
Factor on Factor Regulation

Using the large-scale transcription factor deletion data with a filter imposed such that we only included gene targets that were themselves transcription factors, we converted the raw PCL to a GML format. This ability to visualize datasets from the ArrayPlex Client allowed us to detect pronounced relationships.

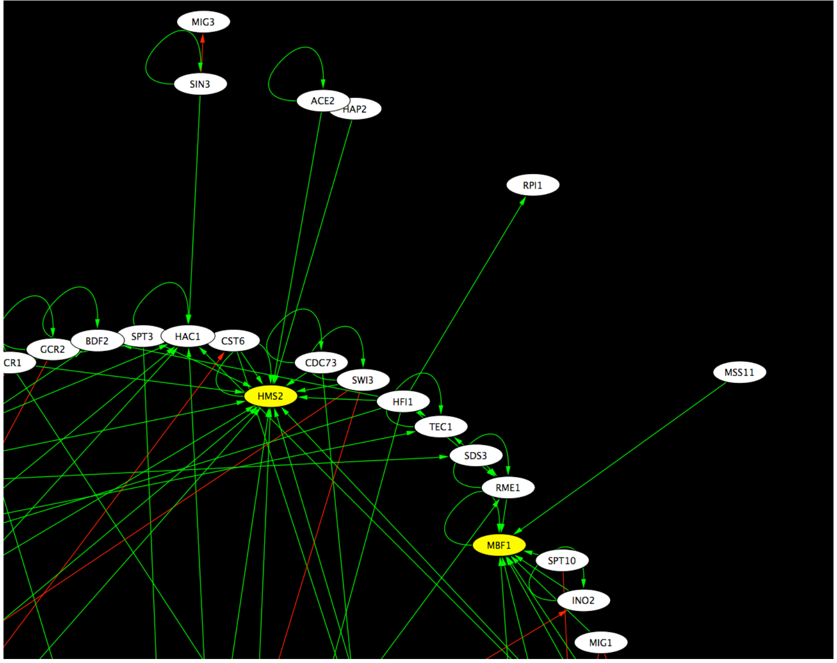
Figure 4.18 (a, b, c, d) depicts this visualization and putative relationships. The transcription factors PHD1, STP4, MCM1, MBF1, and HMS2 each have either a significant count of in-bound or out-bound regulatory connections with the other transcription factors that were profiled. Specifically, MCM1 activates a large number of factors while STP4 is conversely activated by a large number of factors. It is not surprising that MCM1 appears to be an activation hub for many transcription factors in the larger regulatory network. MCM1 has been shown to perform an active role in cell-cycle regulation through regulation of DNA replication initiation [52]. STP4 has little official annotation. The GO ontological enrichment performed on the analysis of its affected targets indicates statistically significant roles in *nucleotidyltransferase*, *polyamine transporter*, *spermine transporter*, and *polyamine* activities. These activities are general to the many pathways of amino acid metabolism and it is thus not surprising that STP4 would then be activated by a wide variety of other transcription factors. Also of interest in the regulatory network, the transcription factors MBF1, PHD1 and HMS2 are each repressed by many factors. Both PHD1 and HMS2 have been shown to perform an active role in pseudohyphal growth adaptation [53, 54]. It is reasonable to believe that their transcriptional abundance would be repressed in the many conditions in which their cellular role is not required.

Figure 4.18 – Factor on Factor Visualization

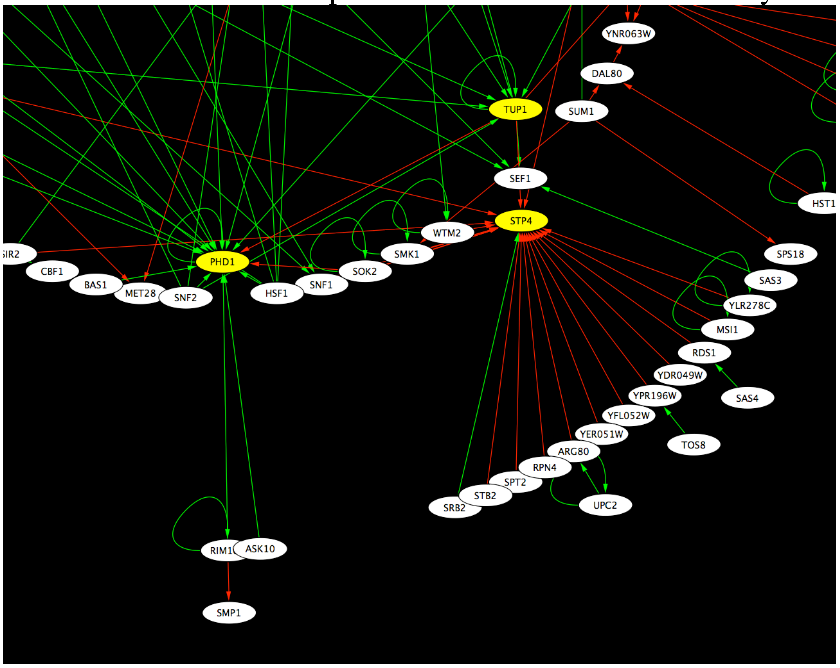
a) The network *G*(*RN*) reduced to regulatory interactions between transcription factors.



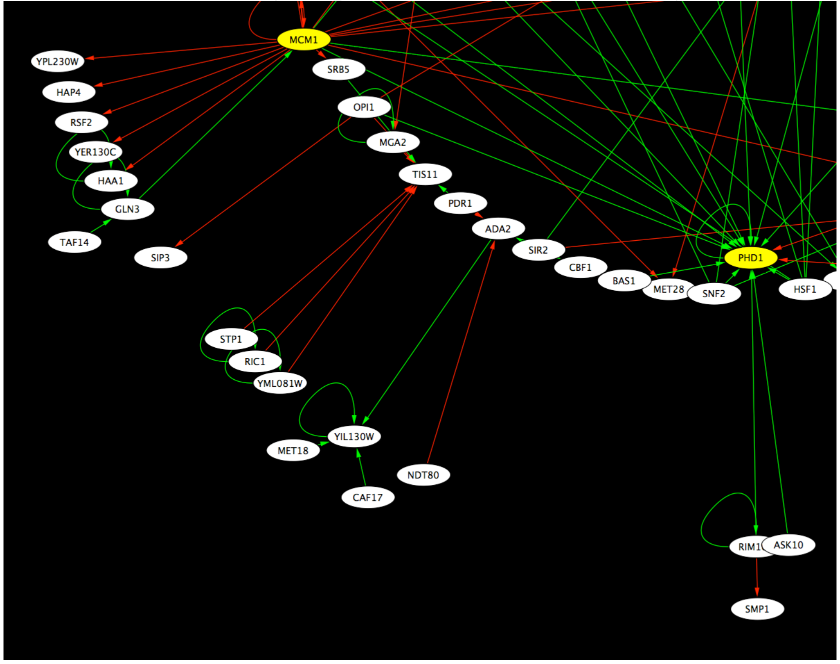
b) The transcription factors HMS2 and MBF1 are each repressed by many factors.



c) The factors PHD1 and TUP1 are repressed while STP4 is activated by many factors.



d) MCM1, characterized in cell-cycle progression, activates many other factors.



DISCUSSION

The results produced by this study are of significant value to the research community. This study represents the largest mRNA expression-based characterization of nearly all *Saccharomyces cerevisiae* transcriptional regulators. The experimental methodology used allowed us to characterize both the relative strength by which each transcription factor regulates each of its KO-regulated targets and the directionality of that regulatory interaction.

Several previous studies have used expression data, graph models, and statistical likelihood of regulation to look at the concept of epitasis within transcriptional networks [55-59]. Our method utilized the inherent strengths of this study to more clearly establish the true set of targets for each transcription factor profiled. The experimental design used growth controls to accentuate transcriptional differences caused by factor deletion and reduce batch-to-batch and other external sources of experimental variation. The measurement of the true transcriptional response of KO-regulated targets provided direct knowledge of whether a transcription factor had an active or repressive role with each target. The implementation of the secondary data model transformation through utilization of the error-model allowed high-confidence targets to emerge from replicate experiments. Our network refinement algorithm used knowledge provided by these design choices, the directionality of factor-to-target regulation and comparable relative statistical significance with which a target was regulated by a factor, to reduce KO-regulated target sets down to a more true set of primary regulatory interactions. It was notable that the process of network refinement was unable to detect non-primary regulatory influences more than four levels deep into the set of regulatory interactions. Of the total unrefined regulatory network $G(N)$, only 1.2% of the regulatory edges were deemed to be indirect and removed from the final network *G*(*RN*). This suggests that the process of regulatory propagation under normal steady-state growth conditions is less prevalent than might have been previously expected.

We analyzed the possibility that RNA binding proteins could exert non-transcriptional effects that would be measured within our experimental process and cause many of the regulated targets to be incorrectly associated with transcription factor deletion. We compared previously characterized targets of RNA binding proteins with the KO-regulated targets of each transcription factor profiled. The expectation was that if RNA binding proteins were playing a secondary role in regulating detected targets we should have found significant overlap between target pools and known targets of RNA binding proteins. No such relationships were detected increasing the confidence that our refined regulatory network was predominately primary relationships.

Similarly we compared all KO-regulated targets sets against each other in a pair-wise fashion. This was done in order to address the concern that target sets could be the result of some non-specific regulation that was the unexpected result of the transcription factor deletion. The process of pair-wise comparing target sets looked for instances in which target sets largely overlapped between two transcription factors even though one factor was not the regulated target of another. We detected no presence of what would be indirect transcriptional regulation by performing these comparisons.

It has been discussed that the process of GO term enrichment uncovered many significant annotations that corresponded with known biological roles which transcription factors have previously been shown to perform. The process of aggregating raw annotation enrichments to composite levels of annotation uncovered and clarified the biological themes behind many of the KO-regulated targets of a transcription factor. Our growth-assay experiments demonstrated that novel functional predictions could be both accurate and testable. One notable problem did occur with respect to the topological analysis of GO term enrichment. We analyzed both the average path length and clustering coefficient of pools of co-enriched terms for a single transcription factor deletion. As previously discussed, these network distance metrics demonstrated that many of our transcription factor deletions resulted in proximally co-located GO term enrichment clusters as compared to random permutations of SGD-assigned term assignments. This conclusion was hindered by the discovery that SGD occasionally assigns proximally close GO terms to the same gene. This meant that a single KO-regulated target could contribute unfairly located GO terms to the pool of ultimately enriched targets thereby skewing the validity of comparison against random permutation. We still believe there are notable untainted results within this network analysis of GO term co-localization. Nonetheless, we did not pursue this line of experimental analysis in the publication of this research.

Our promoter sequence analysis both recovered many previously characterized sequence motifs and discovered a set of novel regulatory motifs for many of the transcription factors profiled. We noted that each of these processes performed very poorly when phylogenetic shadowing was used in place of raw promoter sequence extraction. We had expected use of phylogenetic conservation to increase the signal-to-noise ratio of promoter sequence analysis. Conversely, for both motif search and discovery, phylogenetic shadowing significantly degraded both the quantity and quality of our results. A recent study communicates that cis-regulatory sequences are more mobile than once expected and not well conserved among closely related species [60]. This is one probability that explains our observation.

MATERIALS AND METHODS

Experimental Conditions

Strain Information

All deletion strains were derived from BY4741 (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) and procured from Open Biosystems. The essential transcription factors profiled derive from a BY4741 derivative (*URA3::CMV-tTA MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*). These strains contain a $TetO_{7-}$ promoter that allowed for the conditional repression of transcription factor expression.

Non-Essential Transcription Factor Growth Conditions

Cultures were grown in YPD (1% yeast extract, 2% peptone, 2% dextrose) until mid-log phase and collected for RNA isolation.

Essential Transcription Factor Growth Conditions

The $TetO_{7-}$ promoter of each of these strains was conditionally repressed through the addition of 10 μg/ml doxycycline for 14-16 hours and collected for RNA isolation.

Heat Shock Growth Conditions

Cultures were grown in YPD (1% yeast extract, 2% peptone, 2% dextrose) until mid-log phase and the transitioned from 30°C to 39°C for 15 minutes before rapid collection for RNA isolation.

Overexpression of HSF1

The overexpression strain for HSF1 derives from Y258 (*MATa pep4-3*, *his4-580*, *ura3-53*, *leu2-3*,*112*) and was procured from Open Biosystems. The vector is derived from pRSAB1234 to BG1805. The GAL1 promoter in BG1805 is identical to yeast ChrII bp 278,565-279048. Overnight growth was in SD-ura medium without antibiotics at 30°C. The overnight culture was transitioned to a 25ml -ura 2% raffinose medium overnight. Final growth was a 200ml -ura 2% raffinose medium with a starting OD_{600}. At OD 1.2 the medium was brought to 2% galactose for a 6-hour induction. Cells were the collected for RNA isolation.

Growth Defect Assay

Several transcription factor deletion strains showed functional enrichment for GO terms that directly affect cellular proliferation and survival. These strains were characterized through dilution-assay colony growth under specific growth conditions. Each of the deletion strains *rtg3Δ*, *aft1Δ*, *ric1Δ, pho2Δ,* and *bas1Δ* were spotted on rich-medium (YPD), minimal medium (SD – dextrose, uracil, histidine, methionine, and leucine) and were grown under both normal (30°C) and heat shock (37°C) conditions.

DNA Microarray Methods

Total RNA Isolation

Total RNA samples from the collected cells were extracted by hot acid phenol as described. Cells were resuspended in AE buffer (50 mM sodium acetate, 10 mM EDTA). The cell suspension was mixed with equal amounts of acid phenol (pH 4.5-5.5), and SDS to a final concentration of 0.8 %. The cell suspension was incubated at 65°C for 60 minutes with agitation every 10 minutes. After incubation on ice for 10 minutes, cells were collected by centrifugation. Supernatant was then re-extracted with acid phenol followed by chloroform, then precipitated with ethanol and sodium acetate.

Reverse Transcription

Reverse transcription of was performed using a modified Invitrogen Superscript II protocol (anchored oligo dT 5μg, total RNA 15μg). Amino Allyl dUTP was incorporated into the reverse transcription reaction for the purpose of Amersham Biosciences Cy Dye incorporation. cDNA was purified using Qiagen MinElute columns according to the manufacturer's protocol.

cDNA Fluorophore Labeling

Cy Dye incorporation was performed in the presence of purified cDNA with incorporated Amino Allyl dUTP. Cy Dyes (Cy5, Cy3) were suspended in DMSO and incubated for 60 minutes. Labeled cDNA was separated from unincorporated Cy Dyes through purification with Qiagen MinElute columns according to the manufacturer's protocol.

Microarray Slide Preparation

DNA Microarray slides were post-processed in a solution of succinic anhydride and 1-methyl-2-pyrrolidinone and sodium borate. Slides were plunged rapidly in post-processing solution and agitated for 15 minutes. Slides were then transferred to 95°C water bath and incubated for 90 seconds. Finally, slides were washed with 95 % ethanol and spun dry in a tabletop centrifuge (600 rpm, 3 minutes).

Hybridization & Washing

Hybridization buffer consisted of 50% formamide, 10x SSC, 0.2% SDS.

Purified and labeled cDNA was combined with 2x hybridization buffer and incubated at 42°C for 16 hours. After hybridization slides were washed for 5 minutes in three stages. Stage 1 was composed of 2x SSC, 0.1%SDS (5 minutes). Stage 2 was composed of 1x SSC (5 minutes). Stage 3 was composed of 0.1x SSC (5 minutes). Slides were spun dry in a tabletop centrifuge (600 rpm, 3 minutes).

Computational Methods

Network Modeling

All computational graphs were modeled using the Java package JGraphT, an open-source Java graph library that provides both reusable objects and algorithms for the purpose of modeling, traversing, quantifying, and manipulating mathematical graphs. Significant pre-data prototyping and unit testing was implemented to ensure accuracy of all modeled sub-networks $G(KO^n)$ and the final networks $\mathrm{G}(N)$.

GEO Repository

The expression data from this study has been deposited in the Gene Expression Omnibus (GEO). The series accession number is GSE4654.

Longhorn Array Database

All primary data for every microarray experiment performed in this study is available from the *Publications* section of the Longhorn Array Database hosted by the Iyer Lab of The University of Texas at Austin.

SUPPLEMENTAL INFORMATION

Table 4.5 – Transcription Factors Profiled

The symbol (*) indicates strains where heat-shock was profiled.
The symbol (**) indicates strains with a *tet-off* promoter.

Systematic	**Gene**	**Systematic**	**Gene**	**Systematic**	**Gene**	**Systematic**	**Gene**
YAL051W	OAF1	YER045C	ACA1	YJL103C		YNL021W	HDA1
YBL005W	PDR3	YER051W		YJL110C	GZF3	YNL027W	CRZ1
YBL005W*	PDR3	YER068W	MOT2	YJL127C	SPT10	YNL068C	FKH2
YBL008W	HIR1	YER088C	DOT6	YJL168C	SET2	YNL097C	PHO23
YBL021C	HAP3	YER109C	FLO8	YJL176C	SWI3	YNL139C	RLR1
YBL052C	SAS3	YER111C	SWI4	YJL206C		YNL167C	SKO1
YBL054W		YER130C		YJR060W	CBF1	YNL199C	GCR2
YBL066C	SEF1	YER161C	SPT2	YJR094C	IME1	YNL204C	SPS18
YBL103C	RTG3	YER169W	RPH1	YJR122W	CAF17	YNL216W**	RAP1
YBR033W		YER184C		YJR127C	ZMS1	YNL236W	SIN4
YBR049C**	REB1	YFL021W	GAT1	YJR140C	HIR3	YNL257C	SIP3
YBR083W	TEC1	YFL031W	HAC1	YJR147W	HMS2	YNL309W	STB1
YBR083W*	TEC1	YFL044C		YJR147W*	HMS2	YNL314W	DAL82
YBR103W	SIF2	YFL052W		YKL005C	BYE1	YNL330C	RPD3
YBR150C	TBS1	YFR034C	PHO4	YKL015W	PUT3	YNR010W	CSE2
YBR182C	SMP1	YGL013C	PDR1	YKL020C	SPT23	YNR052C	POP2
YBR195C	MSI1	YGL013C*	PDR1	YKL032C	IXR1	YNR063W	
YBR239C		YGL023C	PIB2	YKL038W	RGT1	YOL004W	SIN3
YBR240C	THI2	YGL025C	PGD1	YKL043W	PHD1	YOL028C	YAP7
YBR245C	ISW1	YGL035C	MIG1	YKL043W*	PHD1	YOL051W	GAL11
YBR275C	RIF1	YGL071W	RCS1	YKL062W	MSN4	YOL067C	RTG1
YBR289W	SNF5	YGL073W*	HSF1	YKL062W*	MSN4	YOL068C	HST1
YBR297W	MAL33	YGL073W**	HSF1	YKL072W	STB6	YOL089C	HAL9
YCL055W	KAR4	YGL096W	TOS8	YKL109W	HAP4	YOL108C	INO4
YCR065W	HCM1	YGL131C	SNT2	YKL112W**	ABF1	YOL116W	MSN1
YCR081W	SRB8	YGL151W	NUT1	YKL185W	ASH1	YOL148C	SPT20
YCR084C	TUP1	YGL162W	SUT1	YKL222C		YOR025W	HST3
YCR106W	RDS1	YGL166W	CUP2	YKR034W	DAL80	YOR028C	CIN5
YDL020C	RPN4	YGL181W	GTS1	YKR036C	CAF4	YOR028C*	CIN5
YDL020C*	RPN4	YGL181W*	GTS1	YKR064W		YOR032C	HMS1
YDL042C	SIR2	YGL197W	MDS3	YKR099W	BAS1	YOR032C*	HMS1
YDL048C	STP4	YGL209W	MIG2	YKR101W	SIR1	YOR038C	HIR2
YDL056W	MBP1	YGL237C	HAP2	YLR013W	GAT3	YOR113W	AZF1
YDL070W	BDF2	YGL244W	RTF1	YLR014C	PPR1	YOR113W*	AZF1
YDL106C	PHO2	YGL254W	FZF1	YLR039C	RIC1	YOR140W	SFL1
YDL170W	UGA3	YGR040W	KSS1	YLR098C	CHA4	YOR162C	YRR1
YDR009W	GAL3	YGR044C	RME1	YLR113W	HOG1	YOR162C*	YRR1
YDR026C		YGR056W	RSC1	YLR131C	ACE2	YOR191W	RIS1
YDR043C	NRG1	YGR063C	SPT4	YLR136C	TIS11	YOR213C	SAS5

YDR049W		YGR067C		YLR176C	RFX1	YOR229W	WTM2
YDR073W	SNF11	YGR089W	NNF2	YLR182W	SWI6	YOR230W	WTM1
YDR123C	INO2	YGR097W	ASK10	YLR228C	ECM22	YOR290C	SNF2
YDR146C	SWI5	YGR104C	SRB5	YLR266C	PDR8	YOR298C-A	MBF1
YDR169C	STB3	YGR249W	MGA1	YLR278C		YOR304W	ISW2
YDR173C	ARG82	YGR249W*	MGA1	YLR357W	RSC2	YOR344C	TYE7
YDR176W	NGG1	YGR288W	MAL13	YLR403W	SFP1	YOR344C*	TYE7
YDR181C	SAS4	YHL009C	YAP3	YLR418C	CDC73	YOR358W	HAP5
YDR191W	HST4	YHL020C	OPI1	YLR442C	SIR3	YOR363C	PIP2
YDR207C	UME6	YHL025W	SNF6	YLR451W	LEU3	YOR380W	RDR1
YDR213W	UPC2	YHL027W	RIM101	YLR453C	RIF2	YPL001W	HAT1
YDR216W	ADR1	YHR006W	STP2	YML007W	YAP1	YPL038W	MET31
YDR216W*	ADR1	YHR041C	SRB2	YML027W	YOX1	YPL042C	SSN3
YDR253C	MET32	YHR124W	NDT80	YML051W	GAL80	YPL049C	DIG1
YDR259C	YAP6	YHR154W	RTT107	YML076C	WAR1	YPL075W**	GCR1
YDR259C*	YAP6	YHR178W	STB5	YML081W		YPL089C	RLM1
YDR266C		YHR206W	SKN7	YML099C	ARG81	YPL089C*	RLM1
YDR277C	MTH1	YIL010W	DOT5	YML102W	CAC2	YPL129W	TAF14
YDR310C	SUM1	YIL036W	CST6	YML113W	DAT1	YPL133C	RDS2
YDR363W	ESC2	YIL038C	NOT3	YMR016C	SOK2	YPL139C	UME1
YDR392W	SPT3	YIL084C	SDS3	YMR019W	STB4	YPL177C	CUP9
YDR421W	ARO80	YIL101C	XBP1	YMR021C	MAC1	YPL202C	AFT2
YDR423C	CAD1	YIL101C*	XBP1	YMR037C	MSN2	YPL230W	
YDR443C	SSN2	YIL119C	RPI1	YMR037C*	MSN2	YPL248C	GAL4
YDR448W	ADA2	YIL128W	MET18	YMR042W	ARG80	YPL254W	HFI1
YDR448W*	ADA2	YIL130W		YMR043W**	MCM1	YPR008W	HAA1
YDR451C	YHP1	YIL131C	FKH1	YMR053C	STB2	YPR009W	SUT2
YDR463W	STP1	YIR017C	MET28	YMR070W	MOT3	YPR018W	RLF2
YDR477W	SNF1	YIR017C*	MET28	YMR070W*	MOT3	YPR022C	
YDR520C		YIR018W	YAP5	YMR075W		YPR054W	SMK1
YEL009C	GCN4	YIR023W	DAL81	YMR164C	MSS11	YPR065W	ROX1
YEL056W	HAT2	YIR033W	MGA2	YMR182C	RGM1	YPR193C	HPA2
YER028C	MIG3	YJL056C	ZAP1	YMR273C	ZDS1	YPR196W	
YER040W	GLN3	YJL089W	SIP4	YMR280C	CAT8	YPR199C	ARR1

Chapter 5: MicroRNA Transcriptional Abundance & Regulation

The research presented in this chapter was begun in May of 2006. Dr. Jian Gu, a graduate of the Iyer Lab, developed our initial process for microRNA expression profiling [61]. Though he was able to attain significant results, the methodology and materials he utilized had significant limitations. I began a process of refinement that eventually led to the development of a new set of procedures for microRNA labeling and hybridization. I have pursued a line of research exploring the regulatory impact the proto-oncogene c-Myc has on microRNA transcriptional abundance. Additionally, I have used serum stimulation of fibroblast cells as a model to further characterize the microRNA component of cellular proliferation. The results of this research are being prepared for publication.

INTRODUCTION

MicroRNAs are members of an extensive family of non-coding RNAs that regulate gene expression in a post-transcriptional, sequence-specific manner [23]. Originally believed to be few in number, limited in biological function, and phylogentically non-conserved, microRNAs have now been identified within nearly all metazoan genomes, including *D. melanogaster*, *C. elegans*, *A. thaliana*, and *H. sapiens*. Contemporary research of microRNAs has primarily focused upon the predictive and experimental identification of target genes and the pathway-specific implications thereof [28, 62-70]. Expression profiling has demonstrated that cellular microRNA quantities can either be temporal or remain constant once initiated. The central focus of this chapter is that microRNA genes, while post-transcriptionally silencing their sequence-specific target, are themselves under some form of active transcriptional regulation.

It has been estimated that c-Myc is involved in the regulation of nearly 15% of all genes [71]. It has been characterized to have a role in many proliferation related cellular processes such as cell-cycle control, apoptosis, cellular differentiation, and DNA damage responses [72]. Previously published results indicated that the oncogene c-Myc is directly involved in promoting the transcription of a specific set of microRNA clusters [39]. We reasoned that the expansive regulation of c-Myc on many gene targets would likely be mirrored in its likelihood to regulate a substantial proportion of microRNA gene transcription. This is supported by the fact that microRNAs have been shown to plan important roles in many of the mentioned biological processes [73, 74]. We theorized that c-Myc modulated microRNAs could be probable post-transcriptional regulatory intermediaries in the control c-Myc has upon its vast set of gene targets.

RESULTS

Experimental Design

The strategy adopted to pursue further understanding of the transcriptional regulation of microRNA genes was built upon two experimental approaches. First, c-Myc is known to be activated by serum stimulation of quiescent cells [61]. We reasoned that serum stimulation would provide an initial biological setting in which we could both recapitulate the previously characterized regulation of the *mir-17/mir-106* microRNA clusters as well as investigate the transcriptional regulation of microRNAs responsive to serum stimulation. Second, direct modulation of c-Myc transcriptional abundance through overexpression and siRNA-mediated knockdown would provide datasets that, when cross-correlated with microRNAs shown to be serum stimulation regulated, would clarify the list of microRNAs most likely to be regulated by c-Myc.

This experimental design was predicated on the capacity to utilize two types of custom-fabricated DNA microarrays to profile the transcriptional abundance of mature microRNAs and mRNAs from matching biological samples. With these datasets the power of computational methods to predict transcriptional regulation and microRNA gene targets becomes strengthened by cellular reality; the true measured state of microRNA and mRNA transcript levels.

MicroRNA Microarray Design

Initial Limitations

Dr. Jian Gu, a graduate of the Iyer Lab, published research detailing the complex physiological response that accompanies cells exposed to various proliferative stimuli [61]. In this work he characterized the global transcriptional response of several fibroblast cell lines including both whole-genome mRNA and microRNA differential expression. This microRNA experiments were performed using the first generation of Iyer Lab microRNA microarrays. These microarrays were functional but had several limitations with respect to their usability and reliability in the study of global changes of mature microRNA expression in mammalian cells.

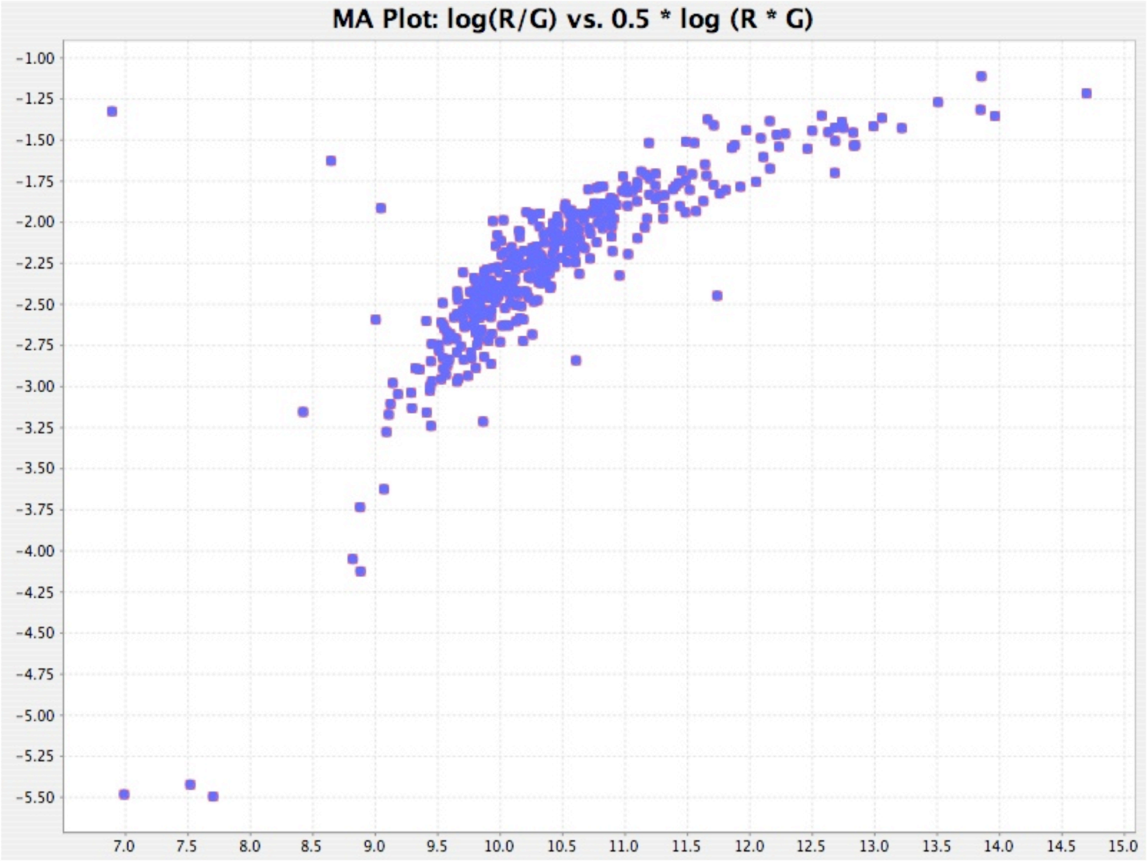
MA-plot analysis of the results produced by these initial microRNA microarrays showed a significant abnormality in the relationship between absolute spot intensities and spot ratios. There should be little relationship between these two spot metrics. Moderate to high intensity spots have no particular bias towards positive or negative log ratio values and produce a distribution of spots that have equal distribution around zero on the log-ratio axis. At low intensities the results are expected to become stochastic yet evenly distributed around zero as the calculation of the log-ratio value becomes unstable. Figure

5.1 depicts a typical MA plot from this first generation of microRNA microarrays. This result is representative of a significant signal-bias observed by three separate research groups at The University of Texas at Austin (Iyer, Harris, and Ellington labs). The visualization shows a skew toward negative log-ratios beginning at moderate and continuing with increased magnitude at low intensity levels.

The consequence of this data abnormality was considerable. Strict data thresholds were required to select for the sub-population of spots not affected by this phenomena. This sub-population was by definition the most intense spots on the microRNA microarray representing the most highly abundant microRNAs in the population of all transcriptionally expressed microRNAs. Data thresholds of this nature would have prevented the characterization of microRNAs expressed at medium to low levels. Secondarily, any log-ratio bias in an experiment severely hampers the capacity to apply normalization algorithms to the primary data. This would have hindered both the believability of any single microarray experiment and relative comparisons one might have made from one microarray experiment to another.

Figure 5.1 – Initial MicroRNA Microarray Signal Bias

The visualization depicts a representative MA-plot from experiments performed on the initial set of microRNA microarrays. A significant signal bias is seen in the form of a skew toward negative log-ratios from moderate to low intensity spots. This signal bias required strict intensity filters to select the small subpopulation of spots minimally affected by the bias. This eliminated the capacity to profile transcriptional abundance for microRNAs expressed at moderate to low levels.



Labeling Methods

Ambion developed the methodology we initially used to label microRNAs. Briefly, mature microRNAs are size-selected and enzymatically extended with a poly(A) tail. These poly(A) tails have the capacity to conjugate fluorophore such that final hybridization yields a signal proportional to the amount of specific microRNA associated with a particular microarray locus. This process functionally worked yet had several shortcomings. First, it was both expensive and time-consuming. The kit required to perform the enzymatic step of poly(A)-extending mature microRNAs doubled the total cost of each hybridization experiment. The protocol for extension was time-consuming and labor intensive, including several dry-downs and lengthy incubations.

Of the many alternative-labeling methods available, the Universal Linkage System (ULS) was a strong candidate. This method involves the non-enzymatic labeling of either DNA or RNA substrates by a reaction that attaches a fluorophore to the N-7 position of all guanine bases. The reaction is complete in less than 15 minutes. The cost per experiment is less than the Cy-dye component of the Ambion-produced poly(A)-extension methodology and eliminated all enzymatic manipulation of the microRNA prior to actual hybridization. Initial testing demonstrated positive results. Overall signal intensities and signal-to-noise ratios measured were significantly higher in ULS-labeled experiments when compared to poly(A)-labeled counterparts.

Slide Substrate

The impetus behind adoption of an alternate labeling method, the signal bias problem, was not resolved through utilization of ULS-mediated labeling of the microRNAs. Several subsequent rounds of hypothesis testing eventually determined that a specific sub-population of the microRNA microarray slides were the true source of the problem. The majority of slides produced were printed on epoxy-coated glass slides purchased from Schott/Nexterion. These slides were expensive but judged to be of higher relative quality. A small minority of the slides produced was from a different supplier. Mock hybridizations without any fluorophore to Schott/Nexterion slides showed these slides produced a green auto-fluorescence at all DNA-spotted positions. Several unsuccessful solutions were attempted in search of a procedure that would make these slides experimentally usable. Ultimately, a new microRNA microarray production run was performed to produce a set of slides usable for reliable experimentation.

Updated MicroRNA Probe Set

Prior to the production of new microRNA microarrays was the acquisition of a new DNA oligonucleotide library that expanded our capacity to characterize a larger set of mammalian microRNAs. This included 281 *H. sapiens* sequences, 49 *M. musculus* variants, 14 *R. norvegicus*, and several dozen negative and positive control sequences. A sample hybridization of this microRNA microarray design is depicted in Figure 5.2.

Development of a Reliable Process

The process of solving the signal bias had several positive related benefits. First was the development of a rigorously detailed process by which microRNA microarray experiments could be reliably performed. This process is depicted in Figure 5.3. Each step in the overall process represents a standardized and verbose protocol that was produced. The protocols ensure that the small RNA population is preserved during each of the experimental sample manipulations: RNA isolation, size-selection, fluorophore labeling, multiple clean-up steps, and final hybridization probe preparation.

Figure 5.2 – ULS-labeled MicroRNA Microarray

Two microRNA samples were differentially labeled with specific ULS fluorophore, combined into a single probe, and allowed to hybridize with complementary spots printed on the surface of the microarray. The data-capture step involves the measurement of relative fluorophore amounts at each spot locus resulting in a visualization composed of red, yellow, green, and black spots.

Red spots represent microRNAs that were more abundant in the experimental RNA sample whereas green spots represent microRNAs more abundant in the reference sample. Yellow spots indicate equal transcriptional abundance between the two samples while black spots represent microRNAs that have low to zero transcriptional abundance in either sample.

The MA plot and population distributions of log-ratio, Cy3 signal-to-noise, and Cy5 signal-to-noise together demonstrate the new microRNA microarray materials and procedures produce high-quality result sets.

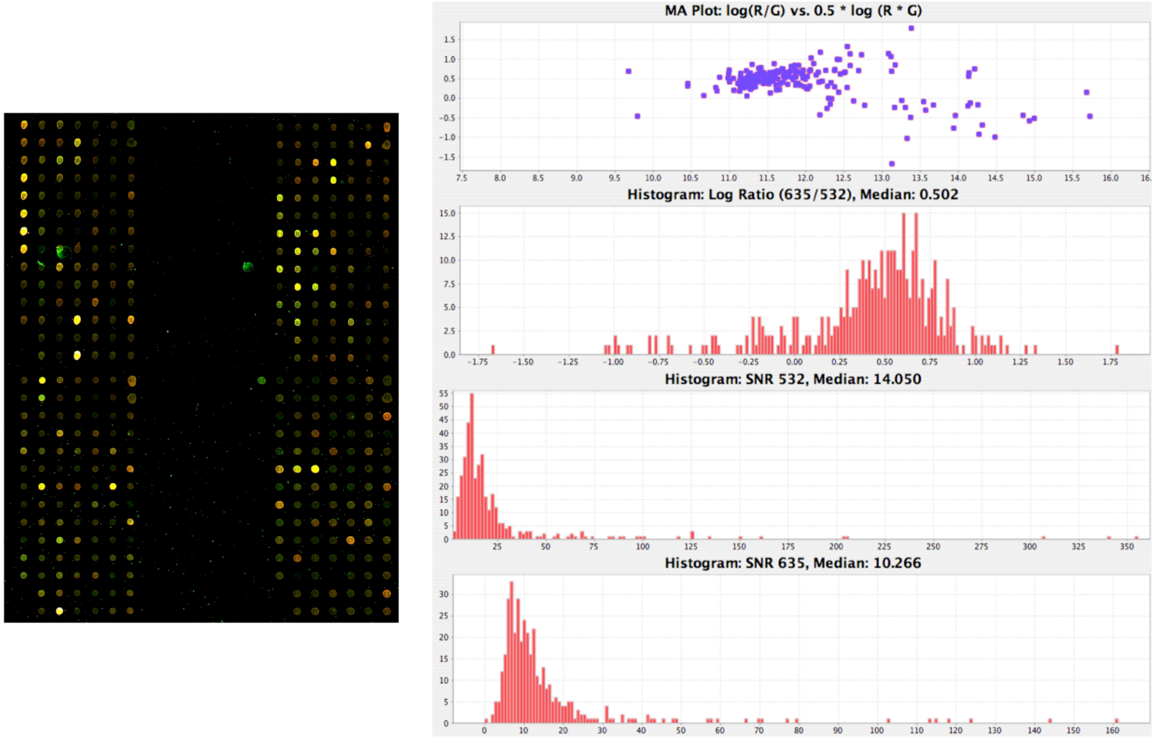
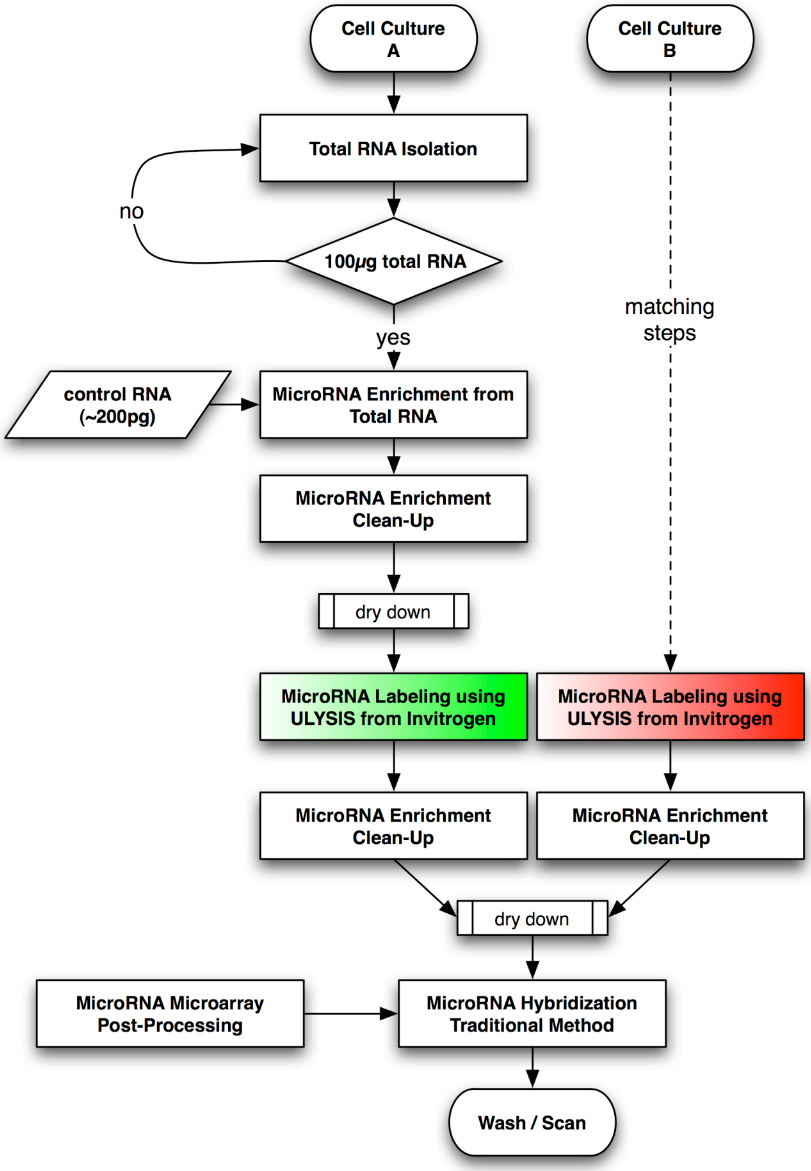


Figure 5.3 – MicroRNA Microarray Procedural Flow

The process by which microRNAs are enriched from cell-cultures for microarray hybridization. Total RNA is isolated using a modification of the Trizol Reagent standard protocol. MicroRNAs are isolated through use of Ambion FlashPAGE poly-acrylamide gel apparatus. Samples are differentially labeled with Invitrogen ULYSIS ULS reagents.


Cell Culture A
Cell Culture B
Total RNA Isolation
no
100μg total RNA
yes
matching steps
control RNA (~200pg)
MicroRNA Enrichment from Total RNA
MicroRNA Enrichment Clean-Up
dry down
MicroRNA Labeling using ULYSIS from Invitrogen
MicroRNA Labeling using ULYSIS from Invitrogen
MicroRNA Enrichment Clean-Up
MicroRNA Enrichment Clean-Up
dry down
MicroRNA Microarray Post-Processing
MicroRNA Hybridization Traditional Method
Wash / Scan

Primary Data Normalization

Two-channel microarray experiments must eventually be normalized [75]. Referring to *MicroRNA Microarray Procedural Flow* depicted in Figure 5.3, it is both possible to introduce sample-specific skew at many of the sub-steps within the overall procedure. Small RNAs can be differentially lost at RNA isolation, size selection, labeling, and clean-up steps. The process of scanning a microarray on two wavelengths typically involves the manual setting of independent laser PMT settings in such a way that maximizes the use of the entire dynamic range on each experimental channel. This can lead to misleading trans-microarray ratio values. Normalization is a process that handles each of the many sources of two-channel microarray data skew. Normalization allows both a single ratio in a single experiment to be believed and for ratio values to be compared across multiple experiments. During the development of the procedures by which microRNA microarray experiments were performed, two strategies were tested.

Positive Control Normalization

Positive control normalization is based upon the process of doping standard amounts of a heterologous nucleic acid into both the experimental and reference sample preparation processes. Nucleic acid loss that occurs during the processing of the experimental or the reference sample can eventually be accounted for if there is a spot or set of spots on the microarray surface that bind these labeled heterologous sequences. If no process-based skew were to occur the outcome would be a log-ratio of zero at positive control spots. Experimental skew where some amount of the positive control nucleic acid was lost in the reference channel would eventually produce positive log-ratios at the positive control spots.

Positive control normalization is a mathematical process that uses a linear coefficient to adjust all spot intensities such that the mean positive control log ratio is zero. The microRNA microarrays were pre-designed with such positive control spots. DNA oligonucleotides combined with T7-mediated transcription of short RNAs allowed for the production of positive control RNA that would both behave much like the experimental microRNAs and bind only to positive control spots on the array. Test hybridizations demonstrated complementarity and specificity between usable RNA and microarray spots.

Global Normalization

In contrast to positive control normalization, global normalization is a more generalized form of experimental normalization and requires no additional experimental input. Global normalization is based upon the assumptions that the average spot is not showing differential expression and that the relative occurrence of repression to activation is proportional. Global mean normalization dictates that all spot ratios are adjusted such that mean log ratio becomes zero. This strategy has been used effectively for thousands of microarray experiments in hundreds of publications.

Normalization Conclusion

Comparison of normalization methods was performed in several ways. Generally, equal mass quantities of positive control RNA were added to both the experimental and reference samples immediately before then enrichment of microRNA from total RNA. The normal microRNA microarray procedures were then followed through primary data capture. At this point both positive control and global normalization coefficients could be determined. The expectation was that the two methods of normalization coefficient determination would yield numerically close values. At the very least it was hoped that they would at least agree upon the overall direction of experimental skew with respect to the global-normalization expectation of population centrality around a log-ratio of zero. The majority of these experimental outcomes fulfilled neither of these expectations. Normalization coefficients determined by the two differential methods were not numerically similar and often disagreed on which way the population of spot ratios should be shifted to achieve experiment normalization.

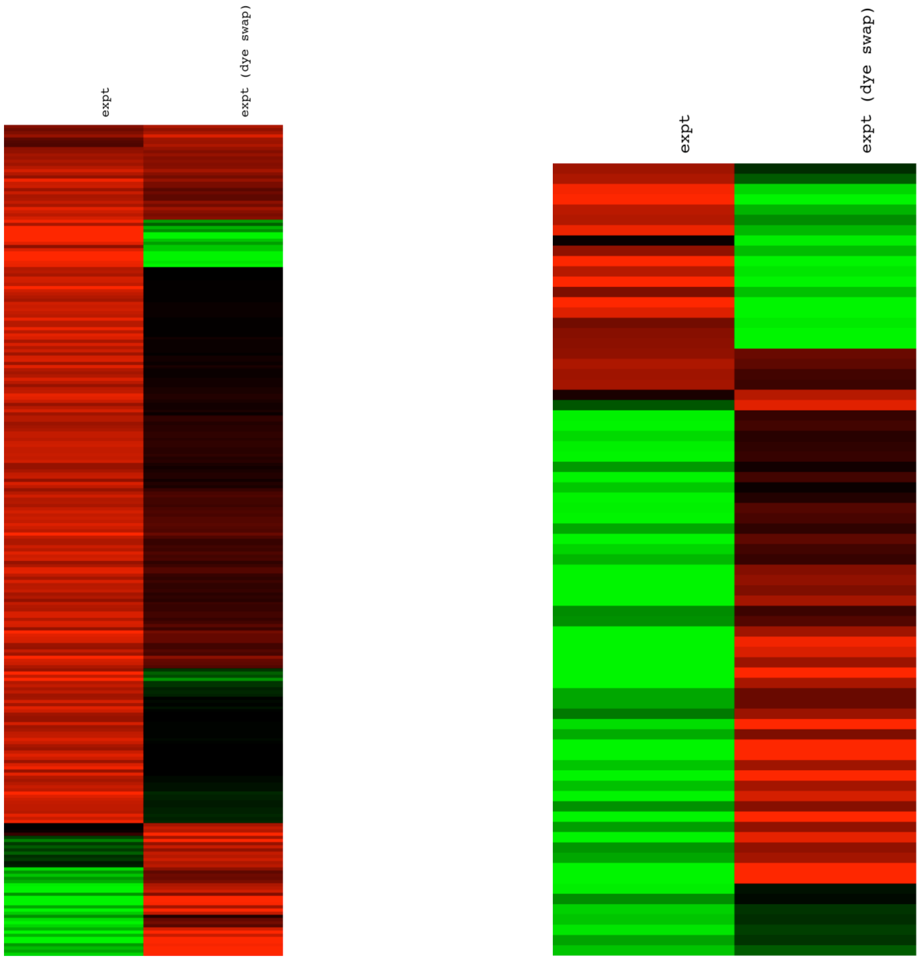
Figure 5.4 demonstrates a typical outcome of an experiment designed to compare the applicability of the two normalization methods. The experiment involved the execution of two microRNA microarray experiments in dye-swap fashion. The null hypothesis of a dye-swap experiment pair is that comparison of post-normalization spot log-ratios from the first hybridization to the dye-swap should show inverse values. A log ratio of 2.0 in the first experiment would be expected to be -2.0 in the corresponding dye-swap. Generally, all red spots in the primary hybridization should match up with green spots in the dye-swap. Positive control normalization consistently produced dye-swap datasets where experiments correlated poorly based upon these expectations. Conversely, global mean normalization reliably produced the expected outcome.

One explanation for the failure of positive control normalization has to do with the ratio of microRNA to total RNA in isolated samples. Positive control normalization is based on the assumption of equal starting masses of both RNA types in a given sample. Global normalization, however, has the intrinsic capacity to handle a situation in which equal masses of total RNA contain significantly different masses of microRNAs. This could be the result of microRNA loss during the RNA isolation procedure.

Figure 5.4 – Normalization Method Comparison

A test of global and positive control normalization was executed in the form of an experiment and dye-swap pair. The expectation of this pair of experiments is that after normalization the experimental and dye-swap log-ratios will be inverse signs.

The data shown is from the same two experiments with the left-pair adjusted by positive control normalization and the right-pair normalized globally. The results show that positive control normalization heavily skews the result set. Conversely, global normalization produces the expected outcome of inverse values microRNA by microRNA basis.



positive control normalization

global normalization

Regulation of MicroRNAs by Serum Stimulation

Experiment Design

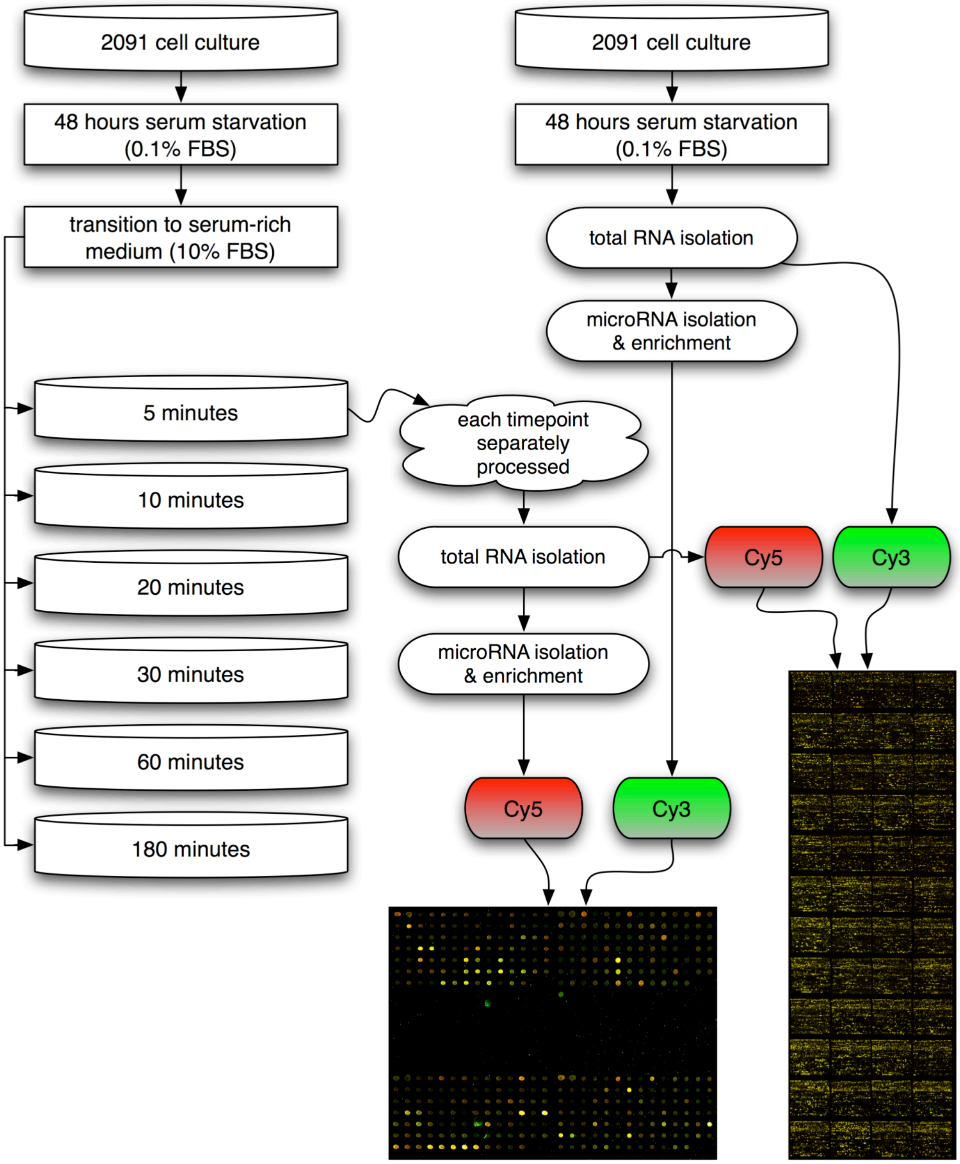
Though an initial characterization has been performed, we believed there was much to be learned about the transcriptional microRNA response in serum stimulated 2091 fibroblast cells [61]. This was especially true in light of the functional limitations of the first generation of microRNA microarrays produced within the Iyer Lab.

The experimental procedure of serum stimulation is detailed in Figure 5.5. Briefly, serum stimulation involved the growth of adherent 2091 fibroblast cells to 40% confluence. Cell cultures were then switched to low-serum 0.1% FBS media for 48 hours. At this time half of the cell cultures were harvested for total RNA. These cultures were quiescent cells and served as the experimental reference. The remaining cultures were serum stimulated by switching them to 10% FBS media. These serum-stimulated cultures were allowed to grow and harvested at specific time intervals of 5, 10, 20, 30, 60, and 180 minutes.

All relative microRNA microarray hybridizations were made between the post-stimulation growth time-points and the reference cultures that were harvested prior to serum stimulation. Additionally, matching mRNA hybridizations were performed with *H. sapiens* DNA microarrays capable of probing for more than 50,000 human ESTs. Each of the time-points was performed in duplicate for both microarray types producing a total of 24 microarray hybridizations. The matching mRNA expression profiling experiments were designed to provide biological validation of predictions between regulated microRNAs and predicted gene targets.

Figure 5.5 – Serum Stimulation Experiment Design

Serum stimulation experiments were performed by 48 hours of serum starvation followed by increasing time intervals of rich medium stimulation. Separate microRNA and mRNA microarray experiments were performed for each time-point (5, 10, 20, 30, 60, 180).


2091 cell culture
48 hours serum starvation (0.1% FBS)
transition to serum-rich medium (10% FBS)
5 minutes
10 minutes
20 minutes
30 minutes
60 minutes
180 minutes
each timepoint separately processed
total RNA isolation
microRNA isolation & enrichment
Cy5
Cy3
2091 cell culture
48 hours serum starvation (0.1% FBS)
total RNA isolation
microRNA isolation & enrichment
Cy5
Cy3
Cy5
Cy3

Baseline Tests

The first analytical focus of the serum stimulation result sets was a relative comparison to previous studies. This approach allowed us to gauge the believability of our novel results by verifying experimental results against independent benchmarks. The results presented in Figures 5.6, 5.7, and 5.8 represent a subset of the microRNAs differentially regulated by our experiments. Concordant with expectations, the microRNAs in Figure 5.6 were activated by serum stimulation [61]. We next cross-correlated the known regulation of the *mir-17*/*mir-106* microRNA clusters by c-Myc [39]. It has been established that one of the many gene targets activated by serum stimulation is c-Myc [61]. It was expected that activation of c-Myc by serum stimulation would produce a cascaded transcriptional activation of these microRNA clusters. This result was confirmed and is depicted in Figure 5.7. Several lines of research have very recently implicated p53 as an activator of *mir-34a* [35, 76-78]. It has been previously demonstrated that in contrast to the expectations of p53 operating as tumor and growth suppressor, p53 is indeed activated by serum stimulation [79]. It was thus expected that *mir-34a* would be activated by serum stimulation. Additionally, the activation of *mir-34* was shown to be specific to *mir-34a* and not expected for *mir-34b* or *mir-34c*. These expectations were confirmed by our result sets and are depicted in Figure 5.8.

Each of these correlated results strengthened our confidence in the analysis of novel microRNA regulatory relationships that emerged from these experiments. Additionally, the specific detection of *mir-34a* demonstrated the specificity of our assay with respect to short, highly related sequences.

Figure 5.6 – Activated MicroRNA Comparison

The microRNAs presented match those previously shown to be activated by serum stimulation [61].

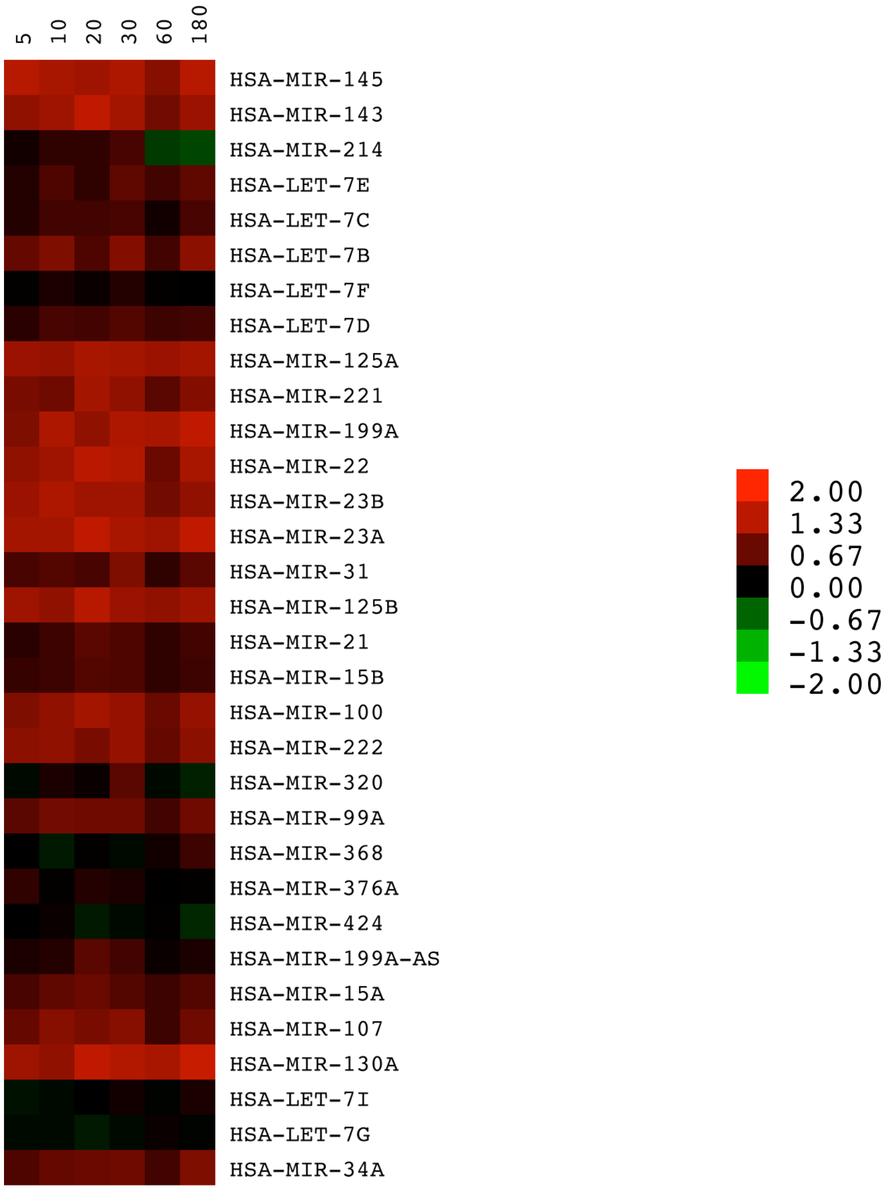


Figure 5.7 – Activation of mir-17/mir-106 Clusters

The transcription factor c-Myc is activated by serum stimulation [61]. Additionally, c-Myc is known to activate the microRNA clusters *mir-17* and *mir-106* [39]. This visualization depicts the relative activation of these clustered microRNAs in our serum stimulation result sets.

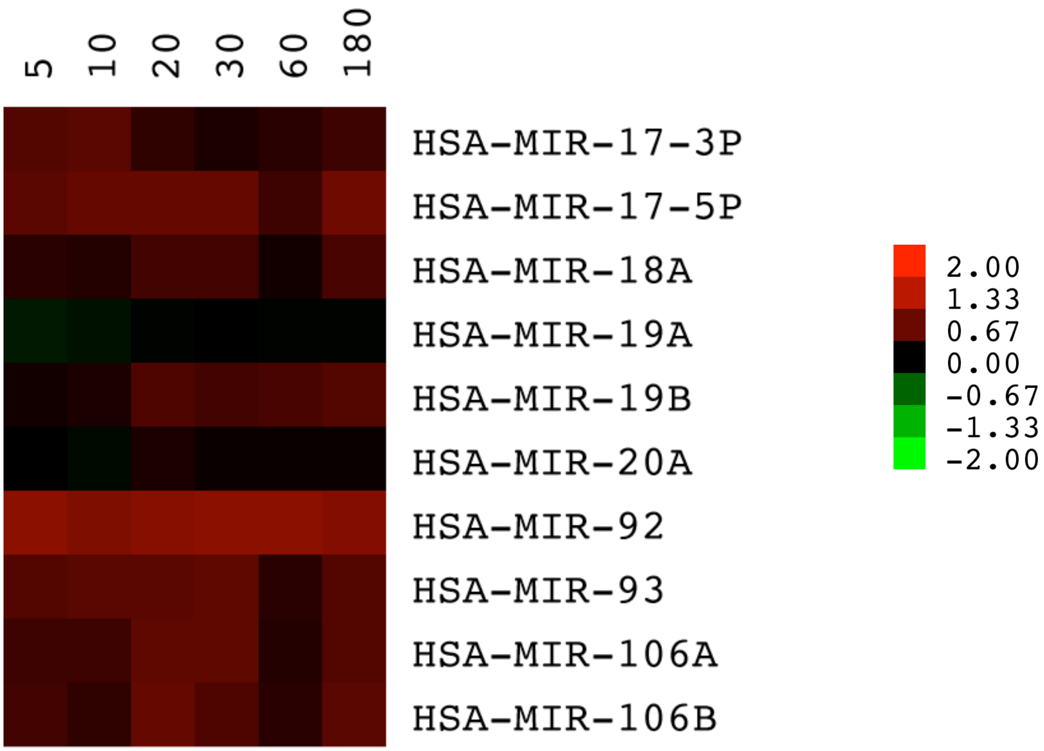
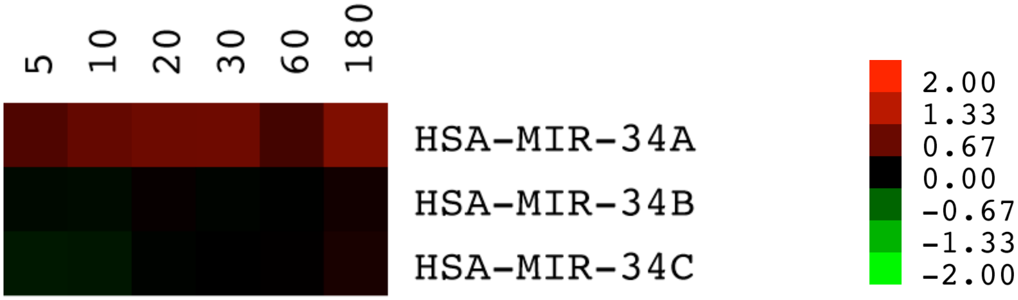


Figure 5.8 – Activation of MicroRNA mir-34a

The tumor suppressor p53 is unexpectedly activated by serum stimulation [79]. Several recent publications have demonstrated the direct activation of *mir-34a* by p53 [35, 77, 78]. Our results demonstrate concordance with these observations by both significance and specificity.



```
CLUSTAL W (1.83) multiple sequence alignment

HSA-MIR-34A      -UGGCAGUGUC-UUAGCUGGUUGUU 23
HSA-MIR-34B      UAGGCAGUGUCAUUAGCUGAUUG-- 23
HSA-MIR-34C      -AGGCAGUGUAGUUAGCUGAUUGC- 23
                   ********  ******* ***
```

Novel MicroRNA Targets

Analysis of the 281 *H. sapiens* microRNA sequences probed for by our microRNA microarrays across the serum stimulation time course produced a set of microRNAs differentially regulated. These microRNA results were extracted from the Longhorn Array Database using absolute intensity filters, spot consistency regression correlation, and log-ratio cutoffs for significance of regulation. All of the microRNA expression profiling presented in this chapter showed differential expression levels that are lower than typically measured by mRNA profiling microarray experiments. For each of the figures in this chapter a color bar is presented to indicate the relative scale of activation or repression measured for a microRNA under a specific experimental condition. These scales are consistently set on a log-ratio range of -2 to +2. This log-ratio transformation is equivalent to a maximum absolute change of 4-fold in either regulatory direction.

Through the use of two log-ratio cutoffs the experimental results for the serum stimulation experiments were partitioned into two overlapping sets. The first, referred to as normal, was based upon the use of an absolute fold cutoff of 1.32 (log ratio of 0.4). The second, referred to as restricted, utilized an absolute fold cutoff of 1.57 (log ratio of 0.65). Appling the aggregation of these filter constraints resulted in a normal and restricted result sets with 145 and 62 significantly activated and repressed microRNAs respectively. These datasets are presented in Figures 5.9 and 5.10. It is notable that the ratio of activations to repressions in the normal dataset is approximately equal while the restricted dataset has a higher relative presence of activated to repressed microRNAs. The analysis process was applied to each of these datasets separately. Unless stated the results presented are based upon analysis of the normal result set.

Figure 5.9 – Serum Stimulation MicroRNA Targets (Normal)

The set of microRNAs differentially regulated by serum stimulation. Results were extracted from the Longhorn Array Database using intensity filters, regression correlation for spot consistency, and an absolute fold-change cutoff of 1.32.

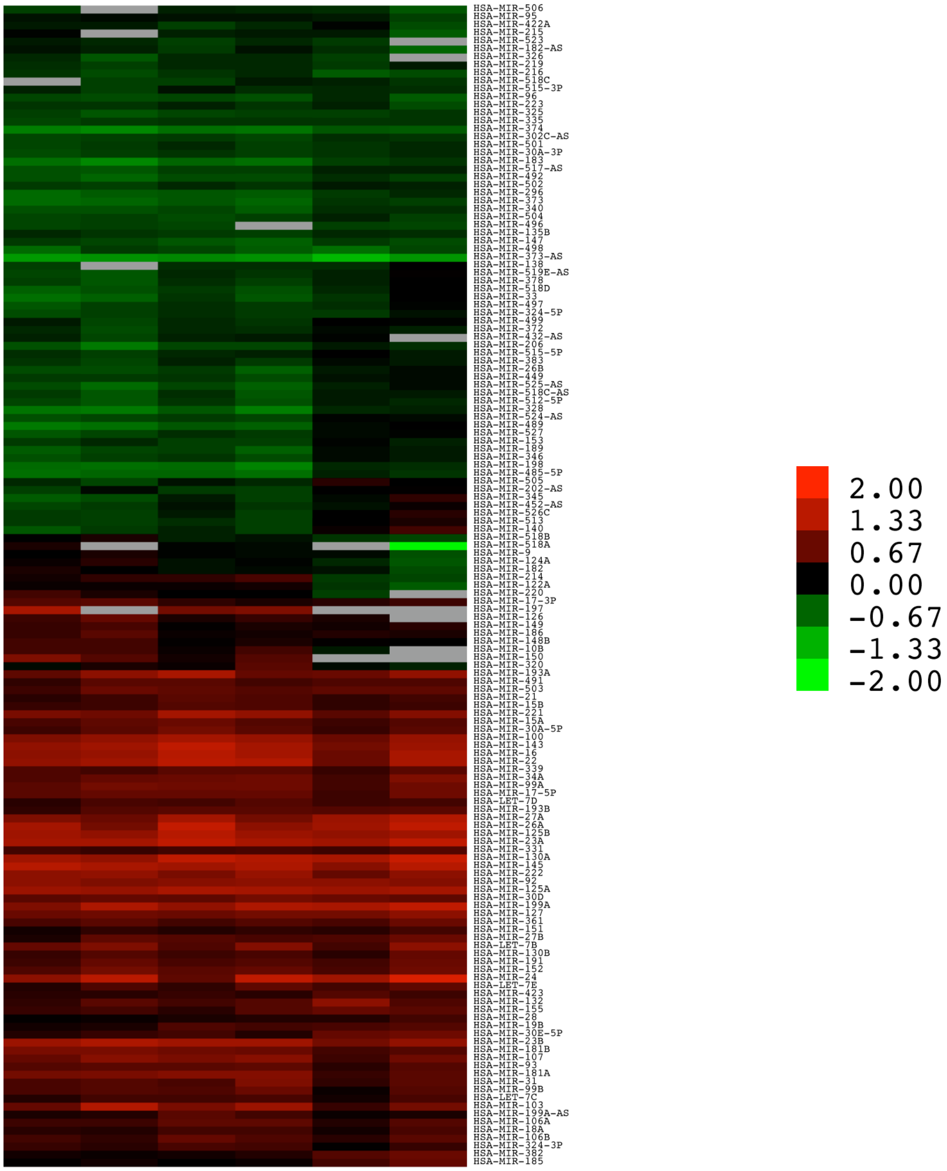
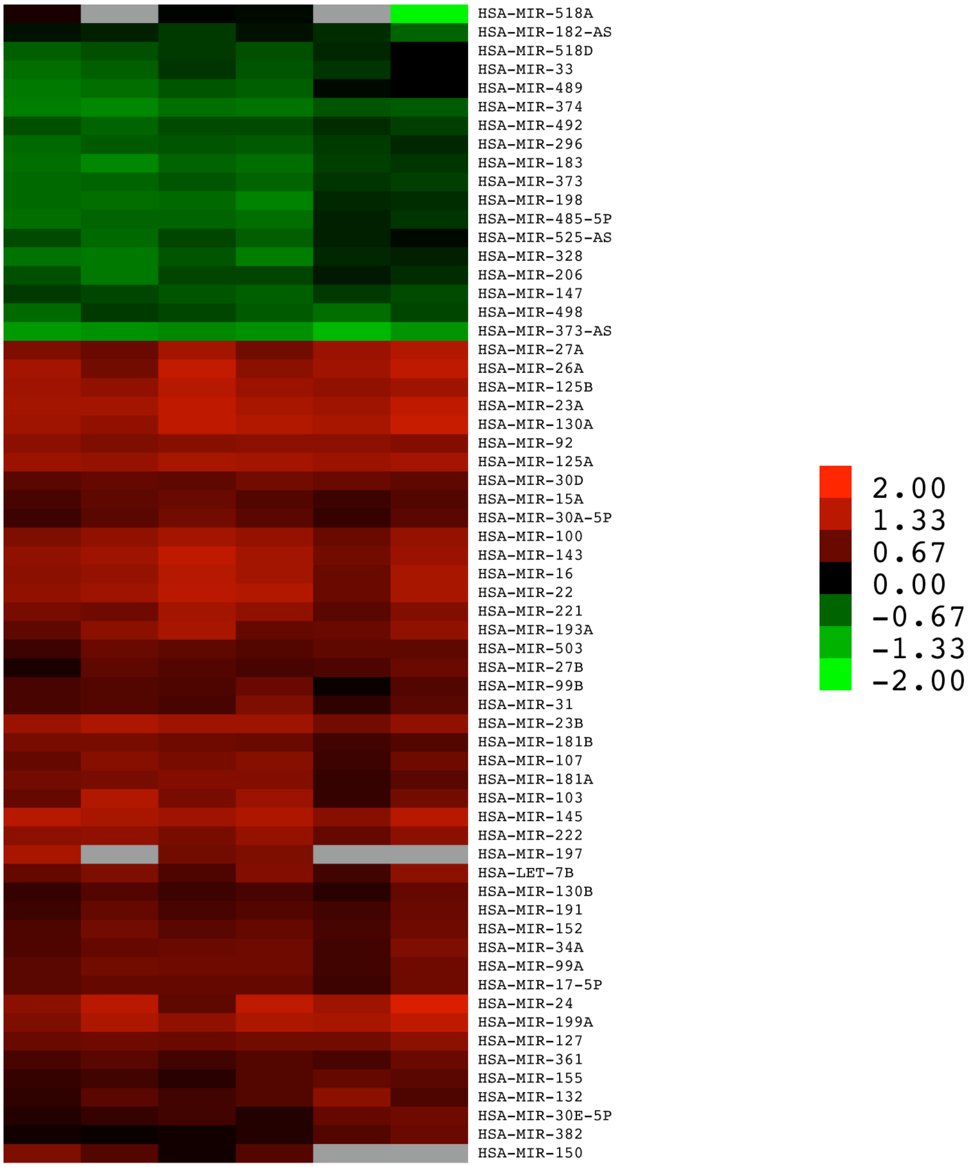


Figure 5.10 – Serum Stimulation MicroRNA Targets (Restricted)

The set of microRNAs differentially regulated by serum stimulation. Results were extracted from the Longhorn Array Database using intensity filters, regression correlation for spot consistency, and an absolute fold-change cutoff of 1.57.



Candidate Regulators

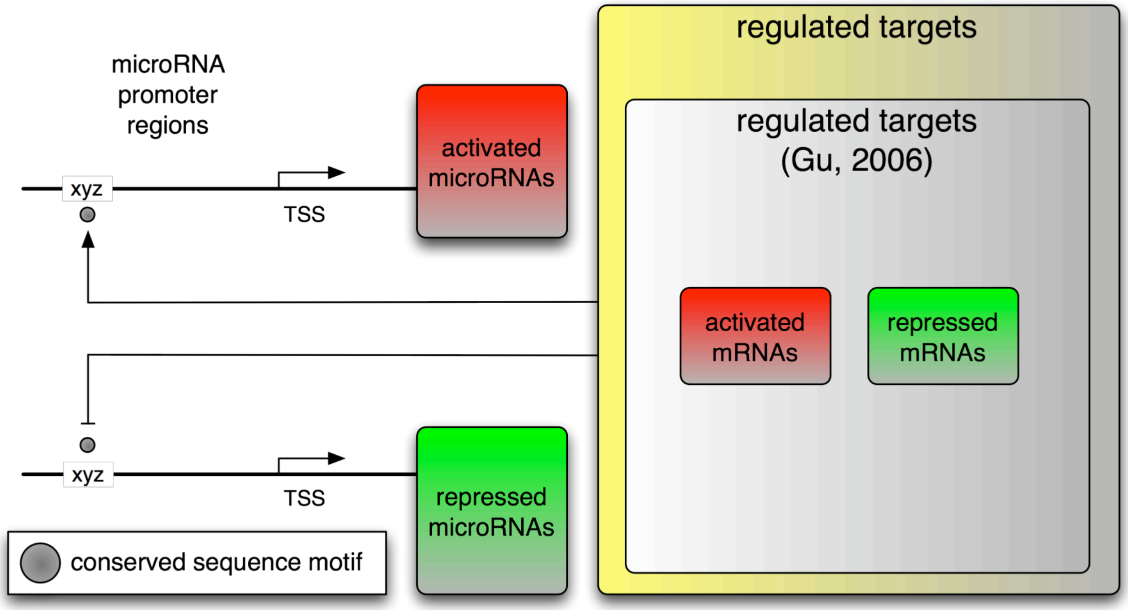
We wished to determine the list of high-confidence transcription factors that were likely to play roles in regulating the set of microRNAs activated or repressed by serum stimulation. Analysis of the whole-genome transcriptional response of fibroblasts under this condition provided an extensive list of factors. This transcriptional response was defined by two data sources. The first was the previously published and referenced 2006 study while the second came directly from the mRNA expression profiling experiments performed with the biological samples used during microRNA expression profiling. We utilized UCSC's predetermined catalog of conserved binding sites for all mammalian transcription factors [TFBS Conserved] to construct a separate candidate list that was representative of conserved sequence motifs within the promoter regions of regulated microRNAs [80]. The mathematical intersection of these separately tabulated lists represented transcription factors that have conserved binding sites upstream of regulated microRNAs and have been shown to be differentially regulated by serum stimulation. We referred to these transcription factors as candidate regulators.

The list of candidate regulators was much smaller than the separate transcription factor lists that contributed to its creation. This list and the strategy by which it was created are presented in Figure 5.11. The transcription factors ARNT, Bach1, Bach2, c-Myc, FOXC1, Nkx3-1, RelA, Sox9, Sp1, SRF, STAT5B, and TGIF were analyzed for ontological enrichment with respect to the aggregation of their collective GO term assignments. No cellular components, biological processes, or molecular functions other than the expected enrichment for transcription-factor associated annotations were detected. Thus, these factors share no common biological theme yet do represent the most likely regulators of serum stimulation mediated microRNA differential expression.

Figure 5.11 – Candidate Regulator Identification

Candidate regulators of microRNAs activated or repressed by serum stimulation were determined through overlap analysis. The set of conserved transcription factor binding motifs present in upstream regions of regulated microRNAs was cross-correlated with both the 2006 study as well as the mRNA expression profiling experiments to determine the list of differentially expressed transcription factors that have a likelihood of binding near regulated microRNAs.

Several of these factors (c-Myc, SRF, Sp1) are known as key regulators in the serum stimulation gene response. Additionally, several factors (FOXC1, Nkx3-1, Sox9, TGIF) have characterized roles in cellular differentiation and tumor suppression.



Factors Identified			
ARNT	c-Myc	RelA	SRF
Bach1	FOXC1	Sox9	STAT5B
Bach2	Nkx3-1	Sp1	TGIF

Candidate Regulator Motif Search

The set of microRNAs regulated were subjected to upstream sequence analysis for statistical enrichment of DNA sequence motifs for each of the candidate regulators. Sequence motifs for each of the candidate regulators were obtained from rVista [81]. Motif search was performed as detailed in Chapter 4 using existing toolsets and the cumulative hypergeometric probability distribution. The search-space was defined as 20kb of upstream sequence relative to the genomic locus at which each mature microRNA is known to reside [82]. Background sequence models included the separate testing of all human microRNA promoters on our microRNAs and all human microRNA promoters.

Significant motifs were determined at a strict *P Value* cutoff of 0.001. Only the candidate transcription factors c-Myc, SRF, and STAT5B had significant motifs.

The canonical c-Myc e-box motif (CACGTG) was not enriched in this analysis. The non-canonical variant (CATGTG) was 75% depleted in the set of microRNAs activated by serum stimulation as compared to background. Depletion of this non-canonical e-box motif was on the basis of total motif over-occurrence to promoters profiled for both background and foreground calculations. The multiple occurrences of e-box motifs within the promoters of these microRNAs were the only qualities that separated them from the average microRNA promoter sequences. This depletion phenomenon is notable and discussed later in this chapter.

The motif V$SRF_Q6 (GNCCAWATAWGGMN) was present in the promoter regions of *hsa-let-7e*, *hsa-mir-125a*, and *hsa-mir-99b*. These three microRNAs are part of a probable but unconfirmed polycistronic cluster [83]. This single shared motif occurrence was evaluated to be significant as it is the only occurrence of the motif in all human microRNA promoter regions.

The STAT5 motif V$STAT_01 (TGCCGGGAA) was present in the promoter regions of *hsa-mir-152*, *hsa-mir-22*, *hsa-mir-23a*, and *hsa-mir-27a*. The last two microRNAs from this list are also members of an unconfirmed polycistronic cluster. The constituents of this cluster have been previously characterized as highly expressed in cholangiocarcinoma growth and proliferation [84].

Candidate regulator motifs were analyzed for over-representation by comparing the ratio of regulated microRNA promoter regions with at least one motif occurrence to all regulated microRNAs with the corresponding background ratio calculation. Under this model the factors Bach2, FOX1C, Sox9, and Sp1 have a ratio at least 20% higher than background. Additionally, the ratio-based comparison was performed comparing the total occurrences of sequence motifs in the pool of regulated microRNA promoters to the total occurrences in corresponding background promoters. The factors Sox9 and Sp1 each had motif concentrations at least 50% greater than background.

Generalized Motif Search

The candidate regulator motif search did not yield a significant set of implicated sequence motifs likely to be mediating the regulation of transcription factors on microRNA gene expression. We next looked at the enrichment of all conserved sequence motifs upstream of regulated microRNAs and compared this enrichment to background expectations.

We partitioned the regulated microRNAs according to whether they were activated or repressed by serum stimulation and extracted 20kb of upstream promoter sequence for each microRNA. For the partitioned microRNA target sets we calculated the occurrence of conserved sequence motifs for all mammalian transcription factors. In order to characterize background occurrence of transcription factor motifs we performed this same calculation using a third collection of microRNAs that represented all human microRNAs present on our microarrays.

The set of activated microRNAs produced a set of sequence motifs that were significantly enriched. The transcription factors Pax-4a, c-Myb, aMEF-2, ZID, Sox5, c-Ets-1, Arnt, and COUP each had enrichment at a *P Value* less than 0.001. The ratio of motif occurrences to microRNA promoters was shown to be approximately 30% above background for each of these transcription factors. It should be noted that while the factors c-Myc, p53, CREB, and SREBP-1a did not have enrichment according to the *P Value* cutoff of 0.001, they each did have sequence motif occurrences that were 20% more abundant than background. Finally, the previously denoted candidate regulators RelA, Sp1, Bach1, Sox9, and Arnt each had statistically insignificant yet relatively low *P Values* and motif occurrences more than 10% above background concentration. It should be noted that the vast majority of mammalian transcription factor motifs profiled in this analysis had motif occurrences well below this 10% threshold.

The set of repressed microRNAs had no enriched sequence motifs. Interestingly, the transcription factors CUTL1, AP-2alpha, E2F, and STAT1 were shown to have motifs significant at the previously stated *P Value* cutoff. These motifs, however, were significant not because of their abundance relative to background but rather their depletion. Each of these motifs has significantly more enrichment in the background of all human microRNAs profiled by our microarrays as compared to the pool of microRNAs repressed by serum stimulation. On average these microRNAs showed 20% less motif concentration when compared to background. The transcription factor CUTL1, the most extreme example, had almost 40% depletion of its motif within the pool of repressed microRNAs. Other transcriptions factors showing similar motif depletion and near the statistical threshold include c-Myb, c-Myc, SRF, and STAT5. Additionally, this set of repressed microRNAs had almost no motifs with ratio-based concentrations greater than background. Conversely, the set of activated microRNAs had no sequence motifs that were depleted in the manner demonstrated by the repressed microRNAs. It is possible that negative regulation of microRNAs includes the loss of cis-regulatory sequences that prevent their activation under specific cellular conditions.

mRNA Targeting

Both the 2006 study and mRNA experiments performed for this study produced a set of genes that were differentially regulated by the proliferative perturbation of serum stimulation. Mechanistically, this regulation could be implemented in many forms. Some genes are primary targets of transcription factors either transcriptionally regulated by the stimuli themselves or post-transcriptionally activated. Other genes may be the result of a regulatory cascade; secondary or tertiary targets regulated by transcription factors that were activated or repressed near the beginning of the initial cellular proliferation signal. Finally, some of these regulated mRNAs are likely the regulatory targets of microRNAs that are differentially expressed.

We wished to characterize which serum stimulation gene targets were likely regulated by microRNA intermediaries. Additionally, we wished to further understand which of the activated or repressed microRNAs were having the largest putative impact on the pool of regulated gene targets.

In order to clarify the answers to these questions we needed one additional piece of regulatory information: the predicted set of gene targets for each microRNA under consideration. This information was provided by the miRBase Targets Database [82].

This database is a compilation of the statistically highest probability gene targets of all characterized microRNAs. It is based upon the combined algorithms of miRanda microRNA targeting, Vienna RNA folding and thermodynamic analysis, and MLAGAN-mediated multiple sequence alignment for the evaluation of conservation of predicted target sequences in the 3' UTRs of candidate gene targets.

Figure 5.12 depicts the information used and operations performed to determine a filtered set of microRNAs likely to regulate gene targets as well as previously known serum stimulation gene targets that are likely the result of microRNA-mediated regulation. Each regulated microRNA was either activated or repressed by serum stimulation. For all activated microRNAs we searched the miRBase Target Database for all putative gene targets. We then correlated this list with a list of genes repressed. Similarly, all repressed microRNAs were mapped to predicted targets that cross-correlated with genes activated by serum stimulation. In this manner we used knowledge of directionality for both regulated microRNAs and mRNAs to correctly associate microRNA regulators with high-confidence gene targets.

The lists of activated and repressed genes derived from two data sources. First the 2006 study was used to determine the results presented in Table 5.1. Of the 73 microRNAs activated by serum stimulation, 71 had predicted targets that showed regulatory repression. Similarly, of the 72 repressed microRNAs, 66 had predicted targets that were activated. These results demonstrate that the majority of regulated microRNAs map to predicted and regulated gene targets. Of 445 total mRNA targets, 127 unique targets were shown to be both measured as repressed and predicted as targets of activated microRNAs. Of the same total mRNA targets, 141 were similarly measured as activated and predicated as targets of repressed microRNAs.

Similar analysis was performed using the second source of activated and repressed target information – the mRNA expression profiling experiments performed on the matching biological samples. Table 5.2 details the outcome of this analysis. The repressed microRNAs mapped to 178 predicted and experimentally activated gene targets. Conversely, the activated microRNAs mapped to 294 predicted and experimentally repressed gene targets.

In order to determine a final high-confidence set of microRNA-mediated gene targets we used the direction-specific overlap of genes regulated by both the 2006 study and the mRNA expression profiling experiments performed as a part of this study. Direction-specific, in this instance, means that for each mRNA activated in the 2006 study we determined a final list of activated gene targets by including only those gene targets that were also significantly activated in our serum stimulation mRNA expression profiles. In a similar fashion we used the direction-specific overlap of repressed gene targets from both datasets to determine a high-confidence list of microRNA-mediated repressed gene targets. These high-confidence results are presented in Table 5.3. Of the original list of microRNA-mediated repressed gene targets presented in Tables 5.1 and 5.2, only 10 were repressed in both mRNA datasets. Similarly, of the microRNA-mediated activated gene targets, only 25 were conserved when both mRNA datasets were used for high-confidence overlap analysis.

The significant reduction in target count through overlap analysis was indicative that the two serum stimulation datasets had significant disagreement with respect to their target sets. These experiments though similar in the biological response they measured, differed significantly in their actual implementation. The 2006 study was performed in such a way that serum stimulated cell cultures were relatively compared to a universal human reference in order to determine targets activated and repressed by the treatment.

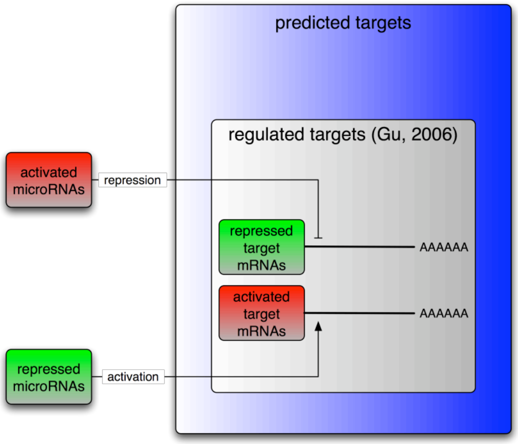
The microRNA and mRNA expression profiling performed as a part of this study was designed to compare the transcriptional response of serum starved cell cultures that were stimulated by transition to rich-medium conditions to reference cell cultures that were similarly starved but not transitioned. In this manner, our experiments measured the time-course response of the cells to a common biological time-point zero. The difference in this design could explain the low overlap of gene targets. Nonetheless, the mapping of microRNAs to predicted gene targets followed by direction-specific mapping to an overlapping high-confidence set of gene targets produced a final set of genes very likely to be regulated by microRNAs during serum stimulation.

Figure 5.12 – Serum Stimulation mRNA Targeting

MicroRNAs were mapped to predicted and regulated gene targets. Directionality of regulation was known for both microRNAs and regulated gene targets. Activated microRNAs were mapped to predicted gene targets that were repressed by serum stimulation. Conversely, repressed microRNAs were mapped to predicted gene targets that were activated.

Figure 5.12a (**right**) demonstrates the method used to generate results in Table 5.1. Gene targets were direction-specific mappings to mRNAs differentially expressed in the 2006 serum stimulation study.

Figure 5.12b (**below**) was used to generate the high-confidence results provided by Table 5.3. Gene targets were direction-specific mappings to the overlap of mRNAs regulated by both the 2006 study and the mRNA expression profiling experiments performed as a part of this study.



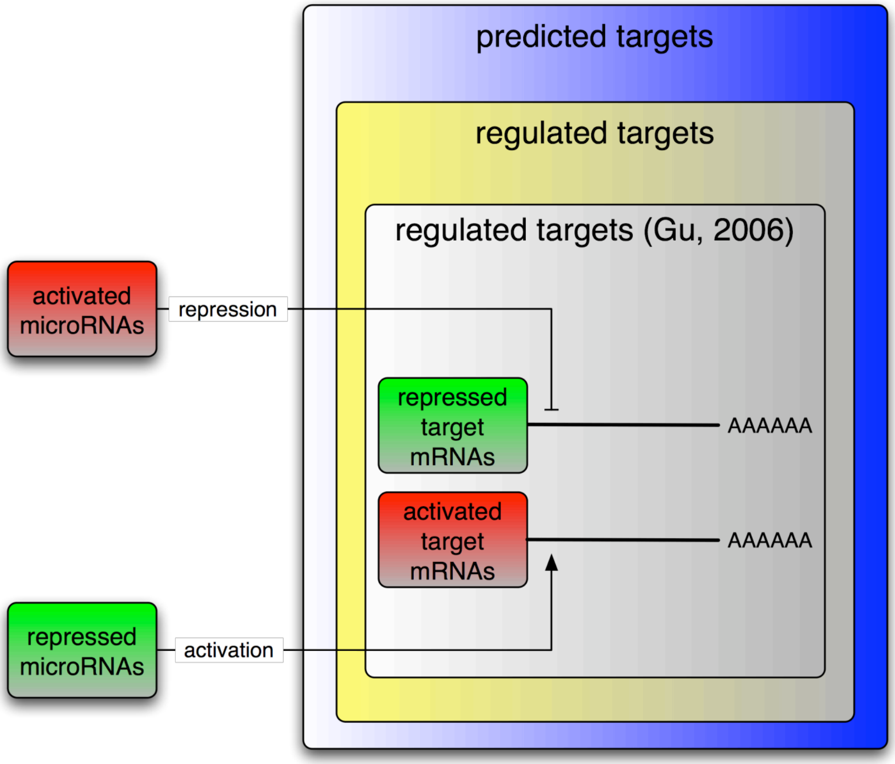


Table 5.1 – Predicted & Known Regulated MicroRNA Targets

Regulated microRNAs were mapped to predicted targets using the miRBase Target Database. Predicted targets were correlated with direction-specific mappings to genes regulated in the 2006 serum stimulation study.

Repressed Targets of Activated MicroRNAs				**Activated Targets of Repressed MicroRNAs**			
ACSL3	FNBP1	MNS1	SLC27A3	ADAMTS1	ESM1	MT2A	SACS
AGPS	FPGT	MRPL34	SLC35A5	AMOTL2	ETS1	MTAP	SDHA
ALDH3A2	FYN	MTMR4	SMARCA3	ANGPTL4	F3	MYC	SERPINB2
APOA2	GAD1	MUT	SMPDL3A	ATF3	FGF7	NAV3	SERPINE1
ATP6AP2	GBAS	MYBL1	SNX2	B3GNT1	FHL2	NEDD9	SERTAD1
ATPAF1	GDF3	NF1	STAT2	BAALC	FHOD1	NFKB1	SERTAD2
AURKB	GIT2	NFE2L1	STX7	BAG3	FOXF1	NFKBIA	SFRS2
BEX1	GPC3	NUDT9	TAF1	BCL10	GADD45B	NOTCH1	SFRS3
BRCA1	GPNMB	NUSAP1	TBL1XR1	BCOR	GATA2	NSUN2	SFTPC
C10orf83	GPRC5B	OGN	TIA1	BDKRB1	GBP1	NUP88	SGK
C18orf10	GRINL1A	PANK1	TNKS	CBFB	GNPNAT1	NUP98	SLC2A14
C9orf126	GRLF1	PARP16	TPD52L1	CCL2	GSPT1	OPRS1	SLC2A3
CABC1	HBLD1	PCNA	TTC19	CCNL1	HAS2	PAWR	SMAD7
CCNG1	HBP1	PLK1	TTC3	CHD1	HIVEP1	PBEF1	SNAI1
CD99	HBZ	POLE3	TUSC2	CKS2	HNRPAB	PDGFA	SOCS3
CDC25C	HEBP1	PPHLN1	UBE2H	CRK	HSPA8	PDLIM5	SOCS5
CDCA1	HK1	PPIG	USP21	CTGF	ID2	PELI1	SPATA6
CDKN1B	HK1	PPP1CC	YEATS2	CYR61	ID3	PFKFB3	SPOCD1
CKAP2	HMGB2	PRDX3	ZFHX4	DDIT4L	IER3	PHC2	SPRY2
COPS4	IGF1R	PRKAG1	ZNF217	DDX21	IL13RA2	PIM1	SPRY4
CRBN	IHPK2	PSAP	ZNF436	DGKH	IL7R	PITPNC1	SRF
CRSP6	ITGB3BP	PSAT1		DKC1	INHBA	PLAT	SSB
CRYZL1	JMJD2A	PTK2		DNAJA1	IRF2BP2	PLAU	STXBP5
DDB2	KBTBD7	PYGL		DOCK10	JARID2	PLEKHJ1	SYN3
DDIT4	KIAA0460	RAB40C		DUSP5	JUNB	PNN	SYNCRIP
EBP	KIF18A	RAD1		EBF	KCNV1	POLS	TNFRSF10D
ECHDC1	KIT	RAD21		EDN1	KHDRBS3	PSG1	TUBB6
EIF2S3	KLHDC2	RBM8A		EGR3	KIAA1949	PTPRO	UAP1
ELMOD2	LETMD1	REV3L		EHD4	KLF4	PURB	UBL3
EPB41L2	MAGED2	SASH1		EIF2C2	KLF6	RAI17	UCK2
EXO1	MAGEF1	SC4MOL		EMP1	KRT18	RBM13	VCAM1
EXOSC8	MAPKAP1	SCAMP2		EMR3	MPV17	RCOR1	VEGF
FBXO25	MARCKS	SCOC		ENAH	MSN	RGS4	VEGFC
FEN1	MDH1	SETDB2		ENTPD7	MT1G	RHOB	WDR1
FGA	MLH1	SH3BGRL		ENTPD7	MT1H	SACM1L	ZNF281
							ZNF347

Table 5.2 – Predicted & Measured Regulated MicroRNA Targets

Similar to Table 5.1. Direction-specific mappings to genes regulated in mRNA expression profiles.

Activated Targets of Repressed MicroRNAs				**Repressed Targets of Activated MicroRNAs**				
ABCG2	COL9A1	KIAA0391	SDF2	AADAT	CRSP6	IGHG1	PAPSS1	SEC61G
ABHD5	CRIP3	KLF6	SDF2L1	ABCB11	CSTF2	IKBKB	PCBP2	SELPLG
ACOT4	DDX58	KRR1	SKIL	ACBD6	CTHRC1	IL22RA1	PCDH17	SERPINE2
ACTC	DENND1A	KTN1	SMAD6	ACIN1	CTSL	ING1	PDZD11	SETD5
ACTN4	DET1	KYNU	SMAD7	ACSL3	DCDC2	INPP4B	PEG3	SF3B5
ACTR3	DKC1	LARP2	SNRPA1	ANAPC5	DCP1B	IQCG	PEX14	SFTPB
AFP	DLGAP1	LATS2	SP100	ANAPC7	DCTD	IRS2	PFDN1	SHB
ALDH3B2	DLL1	LMCD1	SPRR1A	ANKRD11	DDX56	ITM2B	PFDN2	SLC12A9
ANGPTL4	DUSP5	LONRF2	SPTAN1	ANKRD15	DECR2	JAK3	PGC	SLC39A1
ANPEP	DYNLT1	MBNL2	SSTR1	ANXA7	DEPDC2	JOSD1	PHF12	SLC44A1
ARHGAP12	EED	MEF2A	STAT5B	APBA2BP	DES	KARS	PHGDHL1	SLC6A6
ARHGEF11	EGLN2	MEG3	STEAP1	APEX1	DIRAS1	KCNH2	PHTF1	SLIT2
ARHGEF16	EGR2	MGAT1	STRAP	API5	DOCK7	KLF4	PIB5PA	SORCS1
ARL4A	EGR3	MOSC1	SYN2	APP	DPYSL2	KRT5	PLEC1	SPG20
ATF3	EIF2AK2	MT2A	TAOK1	APPBP1	DSG2	LAMA4	PLS3	SSH1
ATP8A2	EMP1	MVD	TBC1D10A	ARIH1	DSG3	LETMD1	PMP22	STAB2
BCL10	ENTPD7	NEDD9	TCL1A	ARPC1A	DTX1	LGTN	POLR2G	STARD3NL
BDKRB2	ERCC1	NFKBIA	TES	ASTN2	DUSP10	LIN7A	POU2AF1	STC1
BHLHB2	EVI5	NOL3	TFG	ATP5B	ENO1	LMAN1	PPIE	STK16
BICD2	F7	NR4A1	TM7SF2	ATP5C1	EPC1	LOXL1	PPIL6	STX16
BMP5	FBN1	PDE1C	TMED5	ATP5G3	EPRS	LOXL4	PPP1CC	SURF1
BPHL	FER1L3	PDE6A	TMEM68	ATP5S	ERGIC3	LRRC50	PPP2R5C	TCEAL8
BST1	FGA	PEG3	TOM1L2	ATP6V0E	ERMAP	LSS	PPP3CA	TCF7L2
C12orf51	FHL2	PGD	TOP2B	ATPIF1	ESCO1	MAF1	PRAF2	TEAD3
C15orf17	FOXC1	PGM2	TPM1	BACH2	ESRRA	MAP3K11	PRCP	TGIF
C16orf45	FOXP2	PHC2	TPM2	BAHCC1	EXOSC2	MAPRE1	PRDX5	TGOLN2
C18orf51	FTSJ3	PIK3C2G	TRPS1	C10orf57	FAM82B	MCM5	PRG2	THBS3
C1orf144	GMCL1	PIM1	TRSPAP1	C11orf10	FAM83D	MDH2	PRMT5	TMEM118
C1QTNF2	GNAS	PKP4	TSPAN2	C11orf74	FER1L3	MDN1	PRR12	TPM4
C3orf19	GNAT1	PLA2G12B	TUBAL3	C4A	FGD2	MEA1	PRR5	TPR
C3orf26	GNB1	PLAT	TUBB6	C6orf115	FN1	METAP2	PSMA1	TPST1
C9orf93	HEG1	PRDM2	TYRP1	C9orf46	FNDC8	MFGE8	PSMA7	TRAPPC4
CAPN3	HIVEP2	PSEN2	UAP1	C9orf75	FXN	MGAT1	PTK7	TRIM38
CASK	HLA-E	PTPLA	UBE2M	CACNA1C	GABARAP	MGST1	PTPRR	TRIP12
CBLB	HNRPA1	PTPN11	UBL3	CACNA2D3	GALC	MGST3	PYCR2	UBAP2L
CCNL1	HNRPR	PTPRM	UQCRQ	CALM2	GAPDH	MICAL1	PYY2	UHMK1
CCT5	HOOK2	PWP1	USP46	CCDC46	GATM	MORF4L2	RAC1	USP14
CDC25C	HS3ST2	RANBP1	WAS	CCDC72	GBAS	MST1	RAD52	USP18
CEACAM5	HSD17B8	RASSF4	WDR37	CD99	GCNT2	MT1H	RAGE	VASN
CEPT1	ID2	RGS4	ZFP36L1	CDC27	GJA5	MTCH1	RBM8A	VEGFC
CES1	IFI30	RHOB	ZNF236	CDC42EP3	GLYAT	MVP	RCN1	VIL2
CES2	IL17RD	SCAMP1	ZNF281	CDK2AP1	GPX3	MYO1B	REC8L1	VTCN1
CIB2	KAL1	SCAP1	ZNF330	CDK5RAP3	GPX4	NAV2	REV3L	WDR22
CKMT2	KCNH2	SDC1		CHCHD5	GRB7	NCF1	RGS10	WEE1
CNKSR2	KIAA0141	SDCCAG1		CHI3L1	GRN	NCOA1	RHOA	WNK1
				CHURC1	GSN	NDEL1	RHOH	WWP2
				CKS2	GSTM1	NDRG2	RP9	YIPF3
				CLN5	H3F3A	NDUFA13	RPL29	YIPF5
				CLYBL	HLA-B	NDUFA2	RPL31	ZBTB2
				CMTM4	HMGN3	NDUFB8	RPL35	ZCCHC10
				CNIH	HNRPH1	NEK6	RPL41	ZDHHC7
				CNKSR1	HSPB6	NFATC3	RPL9	ZDHHC8
				CNN3	HSPB7	NFE2L1	RPS21	ZNF207
				COG2	IARS2	NMNAT1	RPS6	ZNF307
				COL5A1	IDH1	NOC2L	S100A4	ZNF347
				COPS5	IFITM1	NPC1L1	SCAMP2	ZNF498
				COX11	IFITM2	NRP2	SCNN1G	ZNF83
				CPZ	IFRD1	NSMAF	SEC11L1	ZYG11BL
				CROT	IGFBP7	OAZ1	SEC24D	

Table 5.3 – Predicted & Multi-Source Regulated MicroRNA Targets

Regulated microRNAs were mapped to predicted targets using the miRBase Target Database. Predicted targets were correlated with the direction-specific overlap of genes regulated in both the 2006 study and mRNA expression profiling experiments performed.

These results are a subset of those presented in Tables 5.1 and 5.2.

Repressed Targets of Activated MicroRNAs	**Activated Targets of Repressed MicroRNAs**
ACSL3	ANGPTL4
CD99	ATF3
CRSP6	BCL10
GBAS	CCNL1
LETMD1	DKC1
NFE2L1	DUSP5
PPP1CC	EGR3
RBM8A	EMP1
REV3L	ENTPD7
SCAMP2	FHL2
	ID2
	KLF6
	MT2A
	NEDD9
	NFKBIA
	PHC2
	PIM1
	PLAT
	RGS4
	RHOB
	SMAD7
	TUBB6
	UAP1
	UBL3
	ZNF281

Ontology Enrichment for Target Sets

The lists of predicted and coordinately regulated gene targets presented in Table 5.1 were tested for enrichment of GO terms.

The list of repressed gene targets of activated microRNAs were enriched for the GO terms *response to endogenous stimulus*, *DNA repair*, *response to DNA damage stimulus*, *DNA metabolism*, *cell-cycle*, *regulation of progression through the cell cycle*, *anti-oncogene*, *and negative regulation of progression through the cell cycle* at a *P Value* of less than 0.001.

The activated gene targets of repressed microRNAs were enriched for the GO terms *negative regulation of apoptosis*, *negative regulation of programmed cell death*, *apoptosis*, *cell death*, *transcription*, *transcription regulation*, *activator*, *development*, *differentiation*, *growth factor*, *motigen*, *platelet derived growth factor*, *regulation of cell size*, and *positive regulation of cell proliferation* at a *P Value* of less than 0.001. Additionally, the high-confidence list of microRNA-mediated activated gene targets presented in Table 5.2 is exclusively enriched for *regulatory* functions.

It appears that the repressed gene targets are likely targeting cellular processes that would be suspended during cellular proliferation. Similarly, the activated gene targets seem to predominately regulate processes that allow proliferating cells to grow, differentiate, and produce the resources needed for high metabolic and transcriptional activity. The regulatory enrichment produced for activated gene targets in Table 5.2 suggests that microRNAs may have amplified regulatory impact by specifically targeting regulatory proteins for activation.

Regulation of MicroRNAs by c-Myc

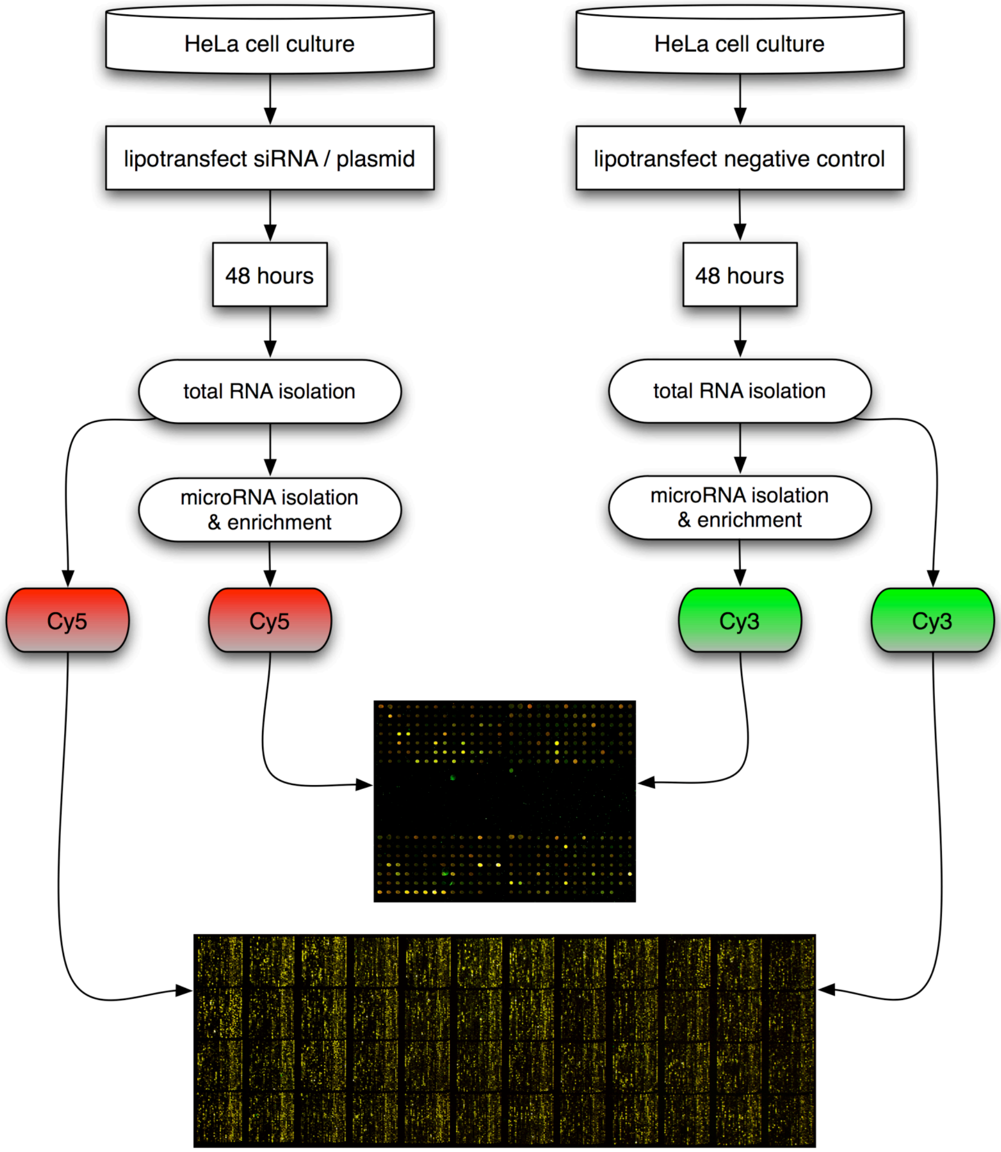
Experiment Design

A 2005 publication established a regulatory link between a well-studied mammalian transcription factor and a specific set of microRNAs [39]. This publication demonstrated that the proto-oncogene c-Myc activates the expression of many constituent microRNAs within three paralogous clusters of human microRNAs. These clusters are referred to as the *mir-17 cluster* (*mir-17a*, *mir-18a, mir-19a*, *mir-20a*, *mir-19b-1*, *mir-92-1*), the *mir-106a* cluster (*mir-106a*, *mir-18b*, *mir-20b*, *mir-19b-2*, *mir-92-2*), and the *mir-106b* cluster (*mir-106b*, *mir-93*, *mir-25*). We hypothesized that over-expression and siRNA-mediated depletion of c-Myc would produce experimental results that biologically validated of our relatively new microRNA microarray process and uncovered novel c-Myc-regulated microRNAs.

Over-expression was performed by lipotransfection of HeLa cells with a c-Myc expression vector acquired from Open Biosystems. Efficiency of transfection was measured using co-transfection of a GFP reporter construct. Western blot analysis of c-Myc verified relative levels of protein abundance between lipotransfected and mock-transfected cell cultures. Similarly, siRNA-mediated mRNA knockdown was performed using synthetic c-Myc-specific RNA purchased from Dharmacon. Similar to the over-expression experiments, HeLa cell cultures were lipotransfected with either the c-Myc-specific siRNA or a mock negative control siRNA. Biological replicates of the over-expression and siRNA-mediated knockdown experiments were performed and both mRNA and microRNA microarrays were separately utilized to capture differential whole-genome expression profiles (Figure 5.13).

Figure 5.13 – c-Myc Overexpression & siRNA Experiment Design

Experimental HeLa cell cultures were lipotransfected with either a c-Myc overexpression construct or siRNA specific to c-Myc. Reference cell cultures were lipotransfected with the appropriate negative control. Total RNA was isolated and used to perform mRNA expression profiling with microarrays. MicroRNAs were isolated, enriched, and were used to perform microRNA expression profiling.


HeLa cell culture
HeLa cell culture
lipotransfect siRNA / plasmid
lipotransfect negative control
48 hours
48 hours
total RNA isolation
total RNA isolation
microRNA isolation & enrichment
microRNA isolation & enrichment
Cy5
Cy5
Cy3
Cy3

Baseline Tests

Similar to the analysis of the serum stimulation set of regulated microRNAs, the initial analytical focus of the c-Myc overexpression and knockdown result sets was a relative comparison to previous studies and expected results. This approach allowed us to gauge the believability of our novel results by verifying experimental results against independent benchmarks.

Concordant with expectations, both the overexpression and knockdown of c-Myc produced respective activation and repression of the *mir-17* and *mir-106* microRNA clusters [39]. This result is depicted in Figure 5.14. Detailed are the results of replicate c-Myc overexpression and knockdown.

In addition to measuring the effect of c-Myc overexpression and siRNA-mediated knockdown on *mir-17* and *mir-106* cluster transcriptional abundance, we realized that the expression profiling of cell-line comparisons could provide additional insight into the differential regulation of these microRNAs by c-Myc. Figure 5.15 depicts the results of this set of experiments. It was previously shown through both mRNA profiling and western analysis performed in the Iyer Lab by both the author and colleagues that HeLa cells have higher mRNA and protein levels of c-Myc when compared to 2091 fibroblast and GM6990 lymphoblastoid cells. We hypothesized that a microRNA-based expression profile comparison of HeLa cells to these cell lines would show relatively higher levels of *mir-17*/*mir-106* cluster expression in HeLa cells. The results of these replicate experiments are presented in Figure 5.15.

Each of these correlated results strengthened our confidence in the analysis of novel microRNA regulatory relationships that emerged from these experiments.

Figure 5.14 – c-Myc Modulation mir-17/mir-106 Cluster Effect

Previous studies have demonstrated that the *mir-17* and *mir-106* clusters of microRNAs are directly regulated by the oncogene c-Myc [39].

Replicate overexpression and knockdown of c-Myc recapitulates this result.

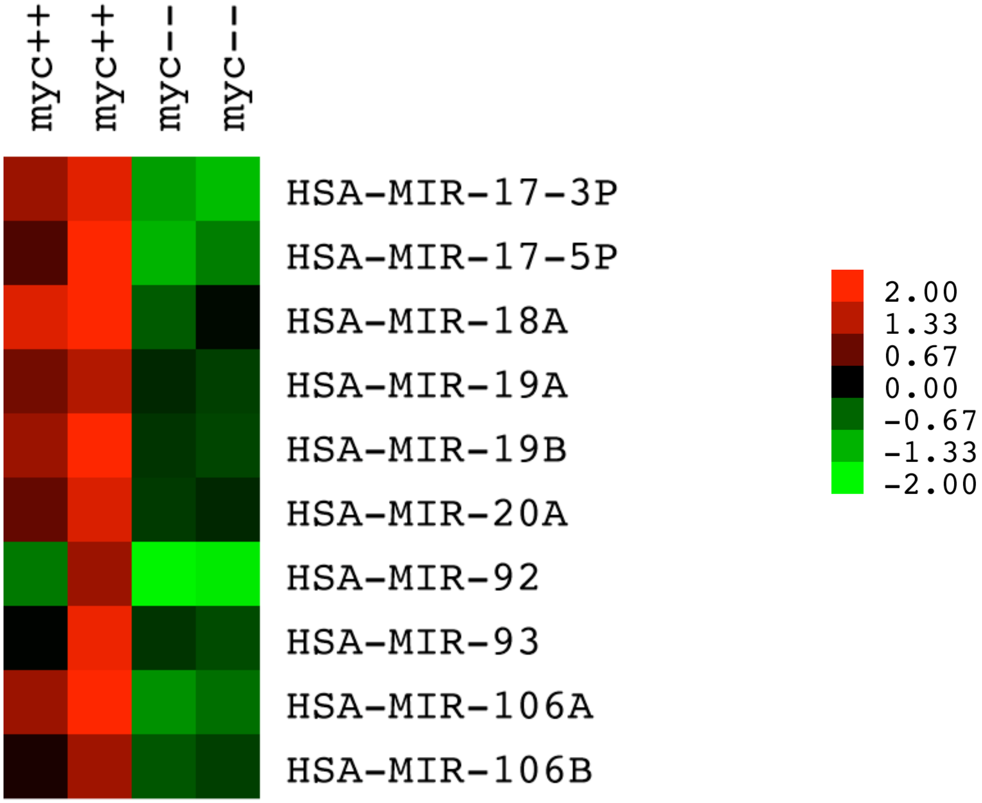
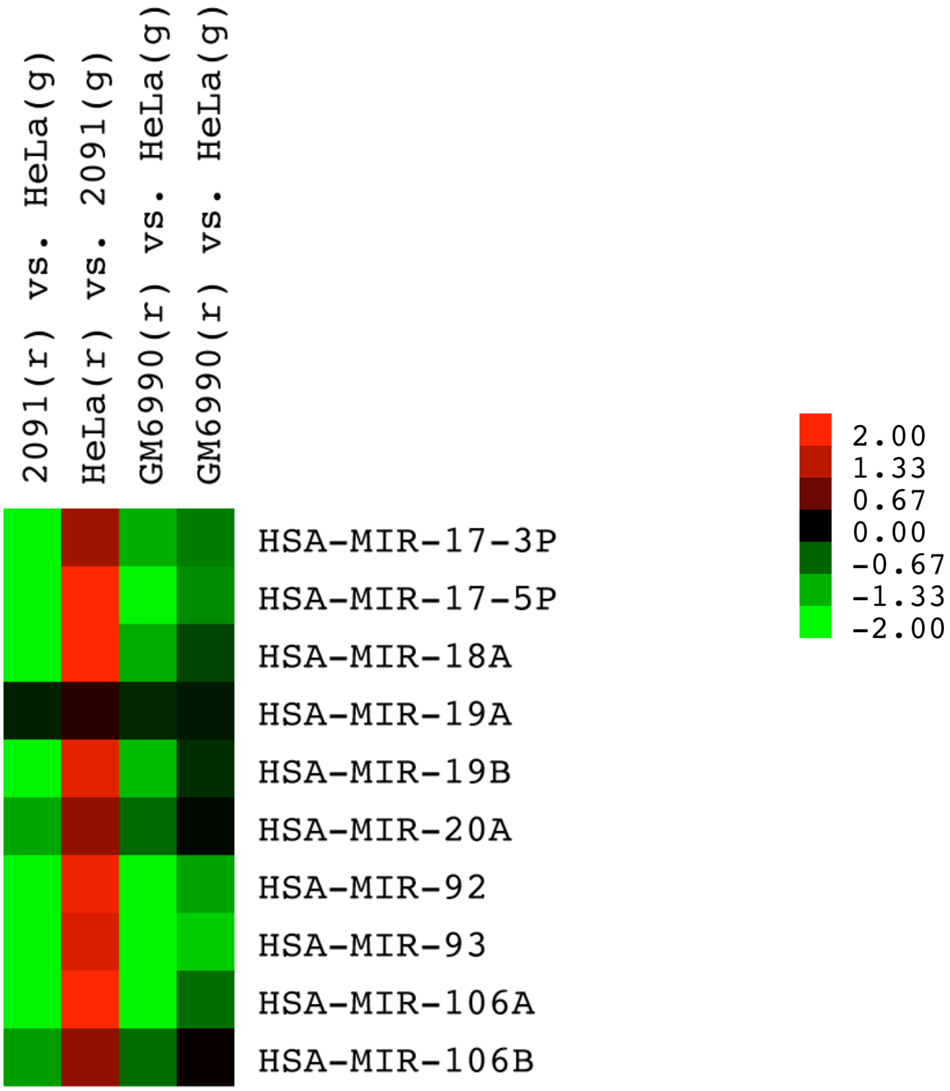


Figure 5.15 – Cell Line mir-17/mir-106 Cluster Effect

MicroRNA expression profiles were performed to compare the transcriptional abundance of *mir-17*/*mir-106* cluster microRNAs in HeLa cells relative to both 2091 fibroblast and GM6990 cells. It was known that the level of c-Myc mRNA and protein was significantly higher in HeLa cells when compared with these two cell lines. Given the known role of c-Myc as a transcriptional activator of these microRNAs, the expectation was that HeLa cells would consistently show relatively higher levels of their expression.



Novel MicroRNA Targets

Several methods were used to determine a high-confidence list of microRNAs regulated by c-Myc. It was hypothesized that alignment of c-Myc overexpression and siRNA-mediated knockdown results would clearly show the opposite directionality of all significantly regulated microRNAs. The behavior of the microRNAs in Figure 5.14 demonstrates an example of this expectation. Activation of microRNAs in overexpression experiments correlates with repression of matching microRNAs in knockdown experiments. Across many replicates of both experiments we observed that while a proportion of regulated microRNAs behaved in this fashion, many did not. We added the results of serum stimulation to the aligned c-Myc overexpression and knockdown experiments to gain a third quantitative perspective.

Figure 5.16 depicts the spectrum of results that were observed. Cross-experiment analysis of c-Myc overexpression and knockdown combined with serum stimulation provided a set of 42 microRNAs activated or repressed under all three experimental conditions. These microRNAs were partitioned into three classes. Class A included the set of 17 microRNAs whose regulatory behavior matched that of our initial expectations. Activation by either c-Myc overexpression or serum stimulation was mirrored by repression under c-Myc knockdown. The 12 Class B microRNAs were those activated by c-Myc overexpression but unexpectedly also activated by c-Myc knockdown and repressed by serum stimulation. Class C is composed of 11 microRNAs repressed by c-Myc overexpression yet both activated and repressed by both c-Myc knockdown and serum stimulation.

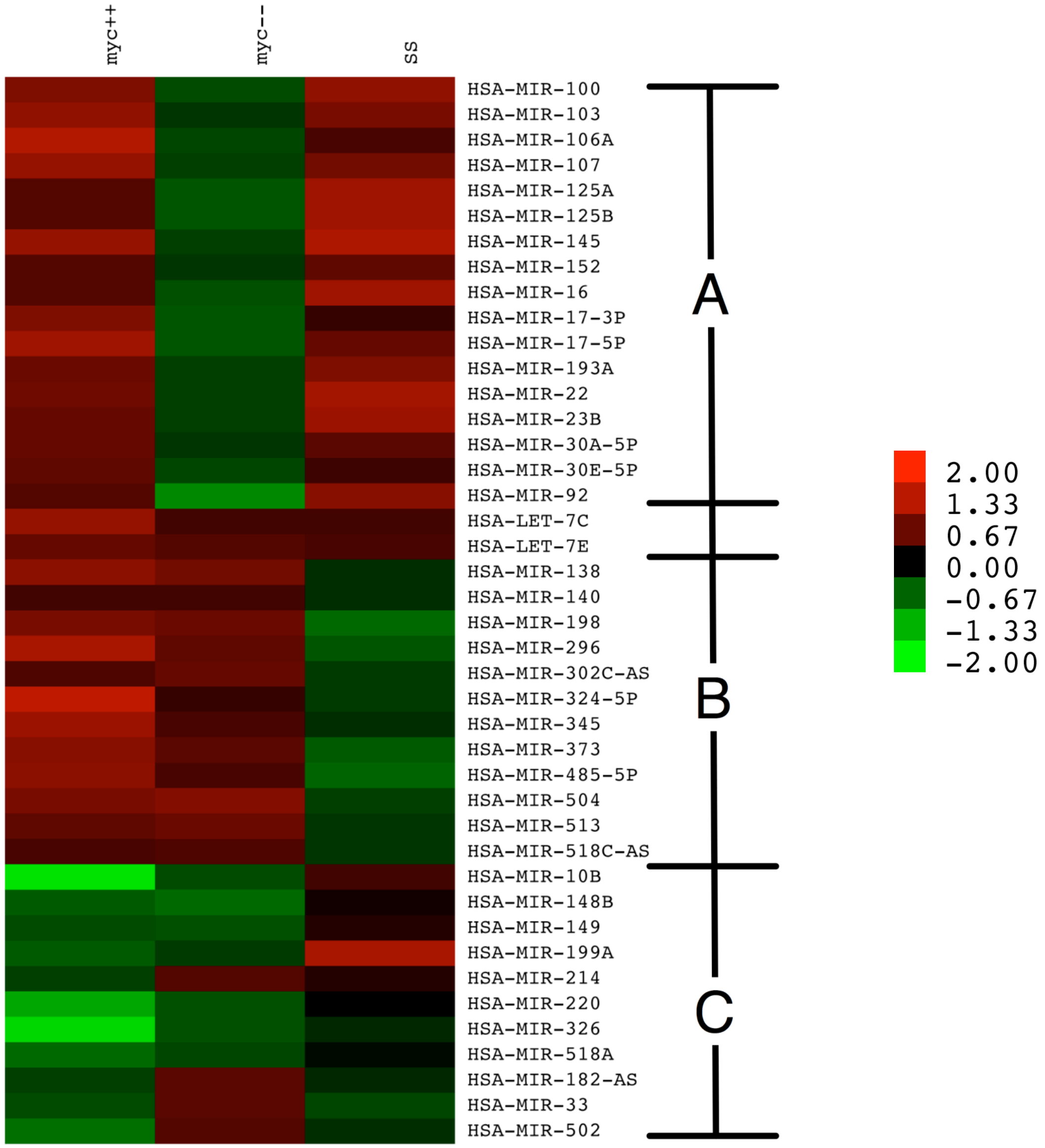
The emergence of these three classes of results was not fully understood. All microRNA microarray experiments were extracted from the Longhorn Array Database such that microRNAs were auto-partitioned by Class A, B, and C membership. Visual inspection of 162 hybridizations produced by both the Iyer and Harris Labs demonstrated no strong tendency for these classes of microRNAs to co-segregate across most experimental conditions. Unsupervised hierarchical clustering, however, did indicate a partial tendency for members of the three classes to co-segregate. This was possibly catalyzed by the experimental results themselves and not necessarily the result of exogenous data bias. For example, Figure 5.16 shows that the c-Myc overexpression and serum stimulation experiments have a strong regulatory signal across each of the three classes. This strong signal, especially when multiplied by correlative biological and technical replicates, had the capacity to direct hierarchical clustering to recapitulate these microRNA class partitions.

To address this question further, the microRNAs in each of the three classes were retrieved for 49 experiments unrelated to c-Myc overexpression, c-Myc knockdown, or serum stimulation. The hypothesis was that other experimental conditions were not likely to create strong regulatory signals on these sets of microRNAs that would mask a tendency for bias in the results. This matrix was hierarchically clustered and inspected for co-segregation along lines of class membership. This procedure showed that the classes significantly dispersed and that their existence was not likely to be the product of some bias intrinsic to the microRNAs or methods used during experimentation.

The *mir-17*/*mir-106* cluster constituents recaptured by our experiments partitioned exclusively to Class A. Eliminating these microRNAs from the 17 members of this class left 13 novel microRNAs that represented the highest confidence microRNAs most likely to be regulated by cMyc: activated by c-Myc overexpression, repressed by siRNA-mediated c-Myc knockdown, and correlatively activated by serum stimulation.

Figure 5.16 – Novel c-Myc MicroRNA Targets

MicroRNAs regulated by c-Myc overexpression (myc++), siRNA-mediated knockdown (myc--), and serum stimulation (ss). The microRNAs partition into three classes based upon their relative activation or repression under the three experimental conditions.



General E-box Search

The transcription factor c-Myc binds the canonical E-box sequence (5′-CACGTG-3′) [85, 86]. It has also been shown to bind non-canonical sequences (5′-CATGTG-3′, 5′-CACGCG-3′, 5′-CATGCG-3′, 5′-CACGAG-3′, 5′-CTCGCG-3′, and 5′-CACGTTG-3′). Each of these sequences was evaluated for enrichment within the pool of Class A, B, and C microRNA promoters. Similar to the sequence motif analysis performed for serum stimulation, 20Kb of upstream promoter sequence was extracted for each microRNA. Enrichment was separately evaluated relative to background promoter collections representing all human microRNAs, all human microRNA on our microarrays, and all human genes.

The canonical E-box CACGTG as well as the non-canonical variants CACGCG, CATGTG, and CACGAG are present in the majority of selected and background microRNA promoter sequences. The other E-box forms are present in approximately 40% of microRNA promoters.

Overrepresentation of E-boxes based on total occurrence within individual promoters showed that the canonical E-box and several of the non-canonical forms occur at high frequencies. None of these E-box motifs were statistically enriched in the selected microRNA promoter regions. Interestingly, however, the Class C microRNAs were repressed by c-Myc overexpression and heavily enriched for over-occurrence of several non-canonical E-boxes. This set of microRNAs showed approximately 100% greater enrichment than background. It is notable that a recent study reported that canonical E-boxes generally correlate with activation of transcription by c-Myc while non-canonical E-boxes are responsible for gene repression [87].

Table 5.4 – General E-box Search Results

E-box motifs were located in all Class A, B, and C microRNA promoters. A recent study characterized canonical motifs as activators and non-canonical variants as repressors [87]. Relative per-promoter and total occurrence rates were evaluated for statistical significance. While the Class A promoters (activated) did not show enrichment for any of the motifs, they did show depletion for a non-canonical form when evaluated for over-occurrence. The Class C microRNAs (repressed) showed significant over-occurrence of several non-canonical variants.

Class A: Single Count Per Promoter							
E-box	Seq	#M	Hit #M	#B	Hit #B	% M	%B
CMYC-C-1	CACGTG	13	11	528	470	84.62%	89.02%
CMYC-NC-1	CATGTG	13	13	528	525	100.00%	99.43%
CMYC-NC-2	CACGCG	13	4	528	201	30.77%	38.07%
CMYC-NC-3	CATGCG	13	7	528	351	53.85%	66.48%
CMYC-NC-4	CACGAG	13	11	528	409	84.62%	77.46%
CMYC-NC-5	CTCGCG	13	5	528	204	38.46%	38.64%
CMYC-NC-6	CACGTTG	13	5	528	220	38.46%	41.67%
Class A: All Occurrences Per Promoter							
E-box	Seq	#M	Hit #M	#B	Hit #B	% M	%B
CMYC-C-1	CACGTG	13	27	528	1507	207.69%	285.42%
CMYC-NC-1	CATGTG	13	74	528	4361	**569.23%**	**825.95%**
CMYC-NC-2	CACGCG	13	7	528	344	53.85%	65.15%
CMYC-NC-3	CATGCG	13	15	528	669	115.38%	126.70%
CMYC-NC-4	CACGAG	13	22	528	876	169.23%	165.91%
CMYC-NC-5	CTCGCG	13	7	528	344	53.85%	65.15%
CMYC-NC-6	CACGTTG	13	7	528	338	53.85%	64.02%
Class C: Single Count Per Promoter							
E-box	Seq	#M	Hit #M	#B	Hit #B	% M	%B
CMYC-C-1	CACGTG	11	10	528	470	90.91%	89.02%
CMYC-NC-1	CATGTG	11	11	528	525	100.00%	99.43%
CMYC-NC-2	CACGCG	11	7	528	201	**63.64%**	**38.07%**
CMYC-NC-3	CATGCG	11	8	528	351	72.73%	66.48%
CMYC-NC-4	CACGAG	11	11	528	409	100.00%	77.46%
CMYC-NC-5	CTCGCG	11	2	528	204	**18.18%**	**38.64%**
CMYC-NC-6	CACGTTG	11	4	528	220	36.36%	41.67%
Class C: All Occurrences Per Promoter							
E-box	Seq	#M	Hit #M	#B	Hit #B	% M	%B
CMYC-C-1	CACGTG	11	32	528	1507	290.91%	285.42%
CMYC-NC-1	CATGTG	11	105	528	4361	**954.55%**	**825.95%**
CMYC-NC-2	CACGCG	11	17	528	344	**154.55%**	**65.15%**
CMYC-NC-3	CATGCG	11	14	528	669	127.27%	126.70%
CMYC-NC-4	CACGAG	11	28	528	876	**254.55%**	**165.91%**
CMYC-NC-5	CTCGCG	11	5	528	344	45.45%	65.15%
CMYC-NC-6	CACGTTG	11	5	528	338	45.45%	64.02%

Extended E-box Analysis

For each of the located E-boxes we decided to evaluate the probability of regulatory binding by c-Myc at each individual genomic locus. The primary goal was to determine the subset of our c-Myc regulated microRNAs whose promoter regions contained the highest likelihood of cis-regulatory capacity.

The first qualitative property that was evaluated for all E-boxes was the degree to which each was phylogenetically conserved across higher eukaryotes. Conservation of non-coding sequences is often a significant indication that evolutionary pressure is maintaining that sequence for the purpose of its cis-regulatory capacity [88]. In addition to the conservation of E-boxes, the genomic binding of c-Myc has been long associated with the proximal presence of hypomethylated genomic regions referred to as CpG islands [89-93]. A recent whole-genome ChIP-PET analysis of c-Myc binding events found that more than 50% of E-boxes that correlated with binding events were within 5kb of a CpG island [90]. This same study noted that c-Myc often interacts with cofactors in order to assert regulatory activation or repression. Finally, one of the most convincing substantiations of the regulatory probability of any cis-regulatory sequence is the experimental detection of actual transcription factor binding.

We proceeded to evaluate all E-boxes of all c-Myc regulated microRNAs based upon these qualities. This extensive process is depicted in Figure 5.17. Conservation was evaluated for each locus through local installation of a portion of UCSC's Genome Browser and command-line execution of UCSC toolsets (hgWiggle) that retrieved conservation metrics for a specified genomic range [80]. The presence of proximal CpG islands was evaluated using the CpG enrichment track [cpgIslandExt]. The existence of proximal cofactor binding sites was evaluated against known genomic locations of all conserved binding sites of all mammalian transcription factors [tfbsConsFactors]. Finally, correlation of located E-box motifs with binding events took advantage of three separate data sources. The whole-genome ChIP-PET results published in 2006 were obtained and associated with our result sets [90]. In addition to these published results, we were able to obtain access to two high-resolution whole-genome datasets that characterized the binding of c-Myc. These datasets, generously shared by Dr. Zheng (Roger) Liu of the Iyer Lab, come from the separate high-resolution tiling microarray hybridization and sequencing of chromatin-immunoprecipitated c-Myc-bound DNA fragments.

The complete results of this analysis are presented in Table 5.5. A total of 4560 E-box motifs were located within 20kb of each human microRNA promoter. For each of the c-Myc regulated microRNAs E-box motifs were evaluated for binding events within 100bp by each of the three whole-genome data sources. Each microRNA with an upstream E-box that correlated with a proximal binding event is listed in Table 5.5. Each of these binding-verified motifs was then evaluated for phylogenetic conservation within the E-box motif itself and for conserved binding sites of other mammalian transcription factors within 100bp. Finally, the presence of CpG islands was searched for on a range suggested by one of the whole-genome binding studies (5kb) [90].

A total of 17 of the regulated microRNAs were shown to have E-boxes with proximal binding events. Of these 17 microRNAs, *hsa-let-7c*, *hsa-mir-152*, *hsa-mir-22*, *hsa-mir-30e*, *hsa-mir-148b*, *hsa-mir-149*, *hsa-mir-199*, and *hsa-mir-214* each have two of the other three qualities shown common to c-Myc regulated targets.

Figure 5.17 – Extended E-box Analysis

E-box motifs located upstream of regulated microRNAs were evaluated for their relative probability of acting as true cis-regulatory sequences. Data sources were gathered to correlate genomic motif locations with phylogenetic conservation across higher eukaryotes, the presence of CpG islands, conserved cofactor binding motifs, and the detection of c-Myc binding in one of three whole-genome binding datasets.

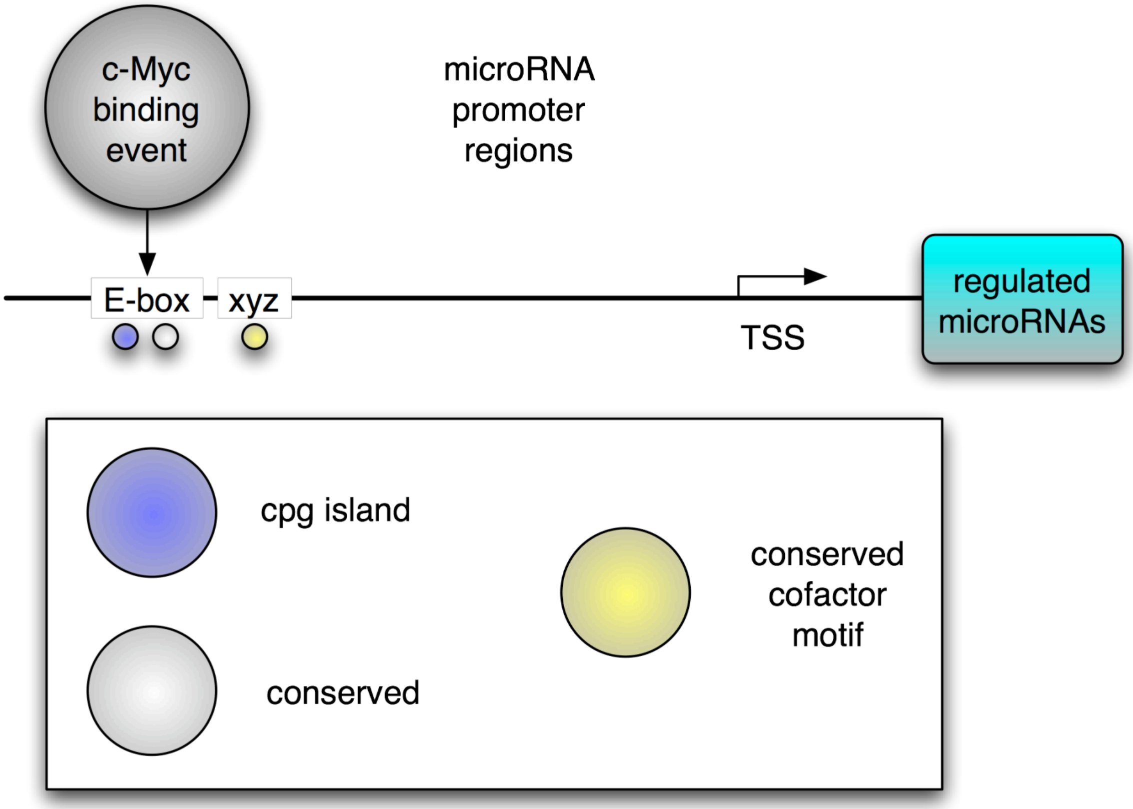


Table 5.5 – Extended E-box Analysis Results

The association of E-boxes in regulated microRNA promoters with information regarding their conservation across higher eukaryotes (cons), the proximal location of cofactor binding sites (co), local CpG islands (cpg), and whole-genome c-Myc binding events (bind). Each record presented had a binding event within 100bp of E-box location. Presence of CpG islands was evaluated by a range of 5kb [90]. The presence of cofactor motifs was evaluated at a threshold of 100bp. Conservation was scored specifically across the E-box motif (Y=near-complete conservation, P=partial).

class	**mirna**	**seq**	**chr**	**str**	**start**	**stop**	**bind**	**co**	**cpg**	**cons**
-	let-7c	CACGTG	chr21	+	16827391	16827396	liu_seq			
-	let-7c	CATGTG	chr21	+	16828826	16828831	zeller_pet	14		Y
-	let-7e	CATGTG	chr19	+	56867959	56867964	zeller_pet			
-	let-7e	CATGTG	chr19	+	56868225	56868230	zeller_pet			
-	let-7e	CATGTG	chr19	+	56868527	56868532	zeller_pet			
A	mir-125a	CATGTG	chr19	+	56868527	56868532	zeller_pet			
A	mir-145	CATGTG	chr5	+	148779464	148779469	zeller_pet			
A	mir-152	CATGTG	chr17	-	43480212	43480217	liu_array	3	1	P
A	mir-22	CACGTG	chr17	-	1574703	1574708	liu_seq	5	2	Y
A	mir-22	CACGTG	chr17	-	1574733	1574738	liu_seq	13	2	Y
A	mir-22	CACGCG	chr17	-	1575313	1575318	liu_array	7	2	Y
A	mir-23b	CATGTG	chr9	+	96881692	96881697	liu_seq			
A	mir-30e	CATGTG	chr1	+	40974711	40974716	zeller_pet	1		P
A	mir-30e	CATGTG	chr1	+	40984270	40984275	zeller_pet			
A	mir-30e	CATGTG	chr1	+	40988656	40988661	liu_array			P
B	mir-138-1	CATGTG	chr3	+	44128096	44128101	zeller_pet			P
B	mir-296	CACGTG	chr20	-	56833512	56833517	liu_seq		1	P
B	mir-485	CACGTG	chr14	+	100571773	100571778	zeller_pet	1		
B	mir-485	CATGTG	chr14	+	100574184	100574189	liu_seq			
C	mir-148b	CACGTG	chr12	+	53004895	53004900	zeller_pet	5		P
C	mir-148b	CACGTG	chr12	+	53007493	53007498	zeller_pet			
C	mir-148b	CATGTG	chr12	+	53010632	53010637	zeller_pet	5		Y
C	mir-148b	CATGCG	chr12	+	53009502	53009507	zeller_pet			
C	mir-148b	CACGAG	chr12	+	53007471	53007476	zeller_pet			
C	mir-148b	CACGAG	chr12	+	53008195	53008200	zeller_pet			
C	mir-148b	CACGAG	chr12	+	53008402	53008407	zeller_pet			
C	mir-149	CACGTG	chr2	+	241025111	241025116	zeller_pet		1	P
C	mir-149	CACGCG	chr2	+	241039855	241039860	liu_array			P
C	mir-182	CATGTG	chr7	-	129212173	129212178	liu_array		2	
C	mir-182	CATGTG	chr7	-	129212173	129212178	liu_array		2	
C	mir-182	CATGTG	chr7	-	129214569	129214574	liu_seq		3	
C	mir-182	CATGTG	chr7	-	129214569	129214574	liu_seq		3	
C	mir-199a-1	CACGTG	chr19	-	10808213	10808218	liu_array			
C	mir-199a-1	CATGTG	chr19	-	10795197	10795202	liu_seq			
C	mir-199a-1	CATGCG	chr19	-	10800823	10800828	liu_seq	4		Y
C	mir-214	CATGCG	chr1	-	170384907	170384912	liu_array	1		P
C	mir-502	CACGTG	chrX	+	49663163	49663168	liu_array			P

c-Myc MicroRNA Intermediaries

Several publications have compiled lists of gene targets thought to be regulated by c-Myc [71, 89, 90, 94]. We wished to characterize how many of these c-Myc-regulated gene targets might be regulated by microRNA intermediaries. We reasoned that some proportion of our binding-verified cMyc regulated microRNAs was acting upon these putative c-Myc gene targets in a manner that fulfilled c-Myc's regulatory role. Similar to the strategy adopted for the serum stimulation gene target analysis, we used direction-specific knowledge of microRNA regulation, miRBase-provided predicted gene targets, and relevant gene target lists to reduce the regulatory permutations to those concordant with experimental results.

Knowledge of c-Myc gene targets was assembled from two sources. First, a multi-source database of cMyc gene targets was recently published [94]. This database has 1737 characterized gene targets but is unable to provide directionality of regulation with respect to whether c-Myc is known to activate or repress cataloged gene targets. For this information we utilized a second source of gene target information. Each of the cMyc overexpression and knockdown experiments were profiled for differential regulation by both microRNA and mRNA microarrays. The mRNA arrays provided a dataset of cMyc regulated gene targets that retains information pertaining to the directionality in which each mRNA was regulated. With this second data source we were able to map between activated microRNAs with predicted gene targets repressed under matching biological conditions. Conversely, we mapped from repressed microRNAs to predicted gene targets that were activated.

The analysis was performed in a manner that separated results obtained using the previously published gene targets from the more filtered results that utilized gene target information produced by our experimentation.

For both sets of analysis, the total number of binding-verified c-Myc regulated microRNAs that were shown to have predicted and c-Myc regulated gene targets was relatively small. Of the microRNAs activated by c-Myc overexpression, *hsa-let-7c*, *hsa-let-7e*, *hsa-mir-22*, *hsa-mir23b*, *hsa-mir-30e*, *hsa-mir-125a*, *hsa-mir-138-1*, *hsa-mir-145*, *hsa-mir-152*, *hsa-mir-296*, and *hsa-mir-485* adhered to this set of criteria. Of the c-Myc repressed microRNAs, *hsa-mir-148b*, *hsa-mir-149*, *hsa-mir-199a*, *hsa-mir-182*, *hsa-mir-214*, and *hsa-mir-502* were compliant with the requirements.

Gene targets were determined by first filtering down to a set of regulated gene targets that were both predicted to be regulated by the set of binding-verified and regulated microRNAs and were regulated in the appropriate direction relative to known microRNA regulation. This list of gene targets was then filtered against the database of curated c-Myc gene targets in order to determine a list of the highest possible confidence.

These lists are succinct and represent a high-confidence assessment of microRNAs likely to be regulated by c-Myc. We were first interested in this line of analysis in order to determine c-Myc gene targets that might be regulated by microRNAs. It should be noted, however, that the mapping of microRNA predicted targets to actual gene targets filters the list of candidate microRNAs to those that have the regulatory potential to act upon mRNA targets known to be concomitantly modulated. Finally, the microRNAs that emerged from this analysis correlate in a near-perfect manner with the subset of regulated and binding-verified microRNAs that possessed E-box-proximal properties that raised their probability of c-Myc regulatory interaction.

Figure 5.18 – cMyc MicroRNA Intermediaries

Regulated microRNAs that with binding-verified E-box motifs were used to search for known c-Myc gene targets whose regulation is likely the result of a microRNA intermediary. Predicted gene targets were mapped for each high-confidence regulated microRNA. These gene targets were filtered against targets regulated by mRNA expression profiling of c-Myc modulation and lists of curated c-Myc gene targets. The resulting set of gene targets represents those with a high likelihood of regulation by microRNA-mediated c-Myc regulation.

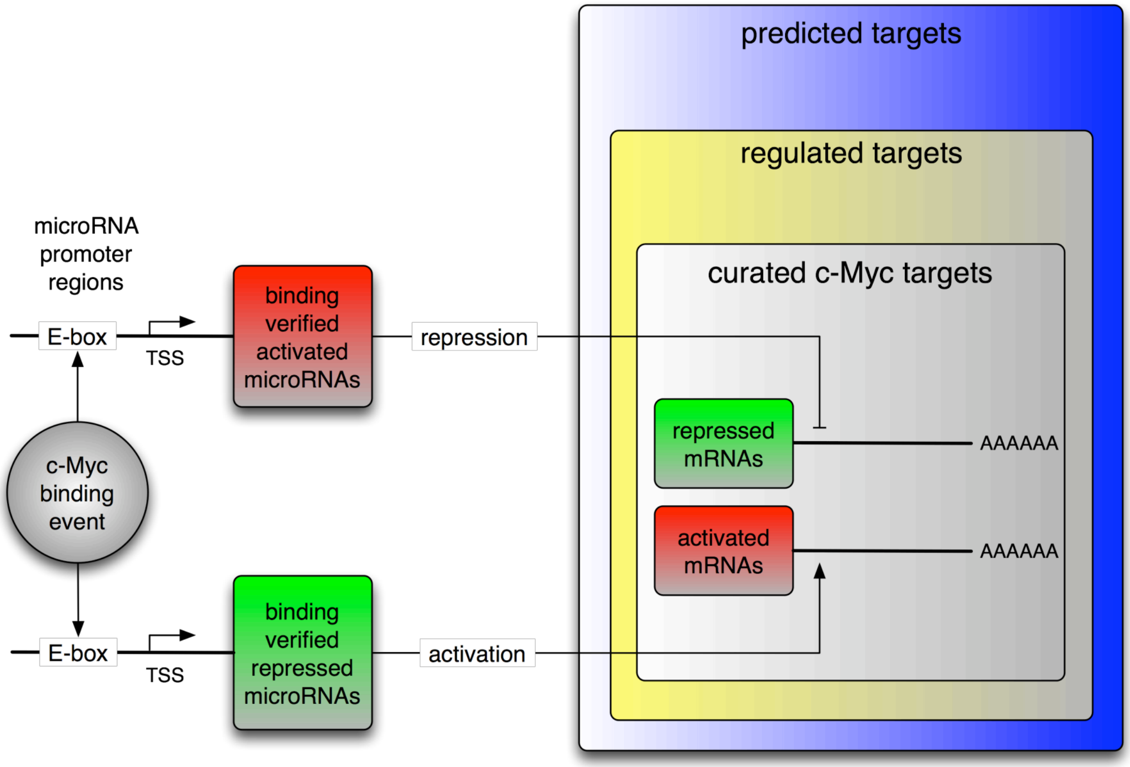


Table 5.6 – Regulated Gene Targets of c-Myc Regulated MicroRNAs

Binding verified regulated microRNAs were mapped to predicted gene targets through use of the miRBase Target Database set of predictions. Predictions were then cross-correlated with direction-specific mRNAs differentially regulated by cMyc overexpression. Finally, the set of gene targets that emerged from this process were filtered by a curated database of known c-Myc targets. The final list provides a set of high-confidence gene targets known to be regulated by c-Myc and likely regulated by c-Myc activated and repressed microRNAs.

Repressed Targets of Activated MicroRNAs				**Activated Targets of Repressed MicroRNAs**
ABHD6	DNAJC8	NHP2L1	STX8	ANXA4
ACBD5	DNM2	PABPN1	TACSTD2	ARF4
ACOT8	DNMT1	PFDN1	TMEM150	BTBD7
ACOX1	DSG2	PHC2	TMPO	C18orf22
ACPL2	ENO1	PLD3	TNFAIP3	CD55
ADSSL1	EPS15	PPAP2C	TPM2	COL3A1
AKR1C2	ERH	PPARA	TPR	DNAJC8
ANXA8	EXOC6	PQLC2	TUBB6	DPM1
AP3S2	FAF1	PRDX1	UBE2D1	EGFR
APOBEC1	FBXW2	PRKAG1	UBE2D3	EIF4A1
APOL1	GAP43	PRMT5	USP39	FN1
ASPHD2	GATM	PSMD14	VEZT	GADD45G
ATG10	GCAT	PTPLAD1	VKORC1L1	HDGF
ATP5F1	GLT8D1	PTPN13	WBSCR22	HES1
ATP9B	GRN	PTPN22	WDR37	HIF1A
BIN2	HDGF	PUS1	WDR71	JUNB
BRCA1	HERC5	PUS7	WISP3	LRP8
BTBD7	HEXIM1	PYGL	WRNIP1	MFAP1
C18orf22	HIP1R	RGL1	ZDHHC7	MRPS7
C18orf22	HLA-E	RHBDD3	ZNF553	MST1
C18orf51	HMGN1	RIC8A	ZNF83	MST1R
C19orf12	HNMT	RIOK1		NIFUN
C1orf128	HSP90AB1	ROR2		PRDX5
C2orf30	IFI30	RPL41		PRPSAP1
C6orf130	IL6R	S100A1		PSMD8
C9orf41	KLHL5	SEC13L1		PUS1
CAPN6	KLK12	SENP5		RPL13
CCNB2	LARP5	SERPINA3		RPL31
CD58	LRP2	SLC31A1		SLC9A3R2
CEP72	MCM3	SLC35E3		SMG5
CLASP2	MDK	SMG5		SNAPC3
CLTB	MFSD4	SNX14		SUPT16H
CMTM7	MNS1	SRFBP1		TFDP1
COPA	MORN1	SSX2IP		TPR
CPXM	MRPL52	ST7L		VAMP3
CWF19L2	MRPS28	STK24		VCAM1
DCBLD1	MYCN	STX17		WDR37
DHX30	NFXL1	STX7		ZNF193

Candidate Cofactors

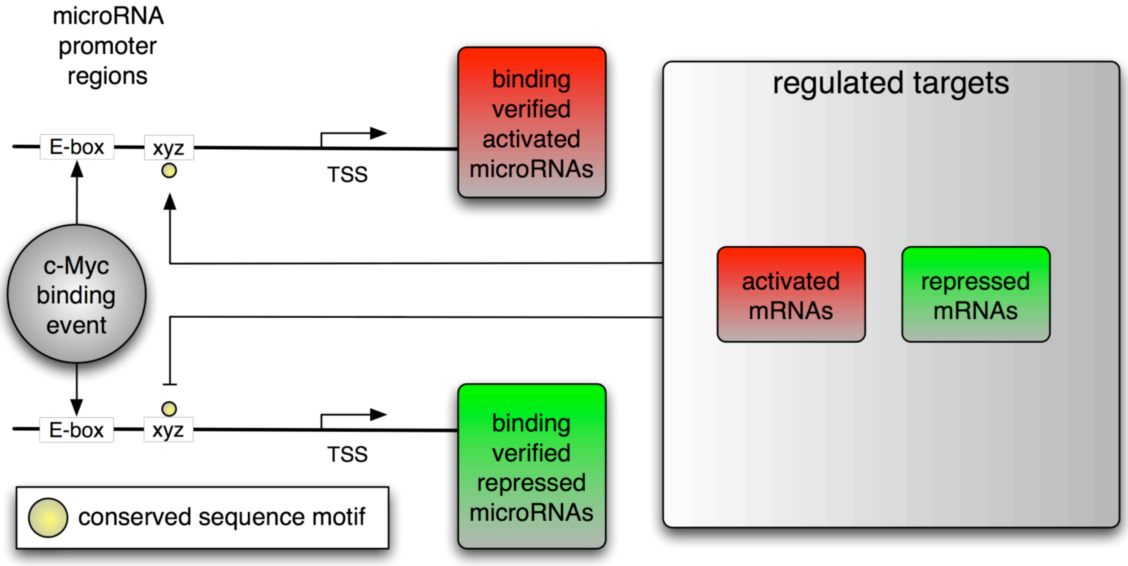
We wished to utilize the knowledge of differentially regulated gene targets to determine candidate cofactors likely to cooperatively interact with c-Myc in the regulation of microRNA transcription. The strategy employed is demonstrated in Figure 5.19. For each binding-verified regulated microRNA we cataloged the list of conserved transcription factor binding sites within its promoter region. The transcription factors represented by these sites were then cross-correlated with transcription factors differentially regulated in the mRNA expression profiling experiments.

The microRNAs *hsa-let-7c*, *hsa-mir-22*, and *hsa-mir-148b* were each shown to have binding sites for the regulated transcription factor Arnt. The microRNA *hsa-mir-148b* was also shown to correlated with STAT1. Finally, the microRNA *hsa-mir-152* was shown to have a conserved binding site for the regulated factor YY1.

Each of these microRNAs was part of the subset of binding-verified regulated microRNAs that had at least two of the qualities that defined a high probability of c-Myc interaction with cis-regulatory sequences. These included phylogenetic conservation of E-boxes, cofactor binding motifs, and the proximal presence of CpG islands. The YY1 motif upstream of *hsa-mir-152* is 42bp from its partially conserved non-canonical E-box motif. Both are located proximal to a CpG island. The microRNA *hsa-let-7c* has a perfectly conserved non-canonical E-box that is 8bp from the cofactor Arnt motif. The promoter of *hsa-mir-22* contains two perfectly conserved non-canonical E-box motifs spaced 30bp apart and proximal to two CpG islands. These two canonical motifs are within 36bp of two separate and conserved Arnt motifs. Finally, the repressed microRNA *hsa-mir-148b* has a fully conserved non-canonical E-box 136bp from its STAT1 motif and a partially conserved non-canonical E-box is 6bp from its Arnt motif.

Figure 5.19 – Candidate Cofactors

For each binding-verified regulated microRNA we cataloged the list of conserved transcription factor binding sites within its promoter region. The transcription factors represented by these sites were then cross-correlated with transcription factors differentially regulated in the mRNA expression profiling experiments to produce a set of candidate cofactors likely to interact with c-Myc in the regulation of microRNA targets.



Real-Time PCR Validation

Our analytical process utilized overlapping and increasingly strict sets of information to arrive at a filtered set of regulated microRNAs. Correlation with both published literature and previously characterized results are an excellent sign of experimental believability but insufficient for final publication. Publication requires an independent experimental assay to probe for and measure the same experimental substrate that is being reported in the primary result. To proactively address this requirement we performed several rounds of qPCR to validate the results of selected microRNA microarray experiments.

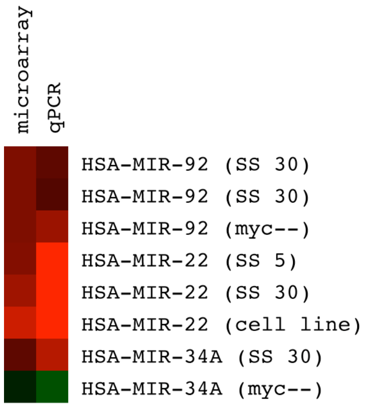
For these procedures we used Applied Biosystems TaqMan microRNA Assays. We specifically selected both reverse transcription primers and qPCR primers/probes for *hsa-mir-92*, *hsa-mir-22*, and *hsa-mir-34a*. Additionally, we selected RPL21 as an endogenous loading control RNA for the purpose of ΔCT and ΔΔCT calculations.

The qPCR results support both the directionality and relative quantity of microRNA activation and repression. Figure 5.20 depicts the candidate set of microRNA and experiments combinations that were used for qPCR validation. For each of these experiments *hsa-mir-92*, *hsa-mir-22*, and *hsa-mir-34a* were relatively profiled using RPL21 as the endogenous loading control measurement. The final comparative results presented were transformed such that they represent comparable same-sign microRNA microarray and qPCR fold-change ratios.

Figure 5.20 – Real Time PCR Validation

Selected microRNAs and experimental samples were subjected to quantitative validation through qPCR. The Applied Biosystems TaqMan MicroRNA Assays were used in conjunction with RPL21 as an endogenous loading control to determine ΔCT and ΔΔCT calculations. Results were normalized to same-sign fold-change ratios and are presented for each microRNA and experiment combination. The experimental samples used are the same samples that were utilized for both microRNA and mRNA expression profiling.

	microarray	**qPCR**
HSA-MIR-92 (SS 30)	1.33	1.23
HSA-MIR-92 (SS 30)	1.33	1.20
HAS-MIR-92 (myc--)	1.33	1.45
HSA-MIR-22 (SS 5)	1.35	2.04
HSA-MIR-22 (SS 30)	1.47	2.86
HSA-MIR-22 (cell line)	1.69	2.13
HSA-MIR-34A (SS 30)	1.23	1.59
HSA-MIR-34A (myc--)	1.08	1.20



DISCUSSION

This study represents a thorough investigation into two dimensions of human microRNA regulatory influence: the transcription factor mediated regulation of microRNA expression and the aggregate regulatory influence of microRNAs on gene targets. The coordinated expression profiling of biologically matched mRNA and microRNAs samples under a spectrum of physiological perturbations allowed for cross-correlative analysis of post-transcriptional microRNA regulation of gene targets as well as reciprocal gene-mediated regulation of microRNAs. Each phase of analysis began with significantly long lists of putatively significant microRNAs and transcription factors and used processes of cumulative evidence correlation to determine succinct lists of high-confidence regulators.

Process Determination

Many sections of this chapter began with efforts to correlate experimental results with literature-based expectations. These efforts were required to gain confidence in what was a relatively new experimental system for the research environment in which they were performed. Each example of correlative substantiation demonstrated that our experimental process was capable of fidelity with respect to the capacity to detect and report microRNA expression profiles across a spectrum of transcriptional abundance. Additionally, the specific detection of *hsa-mir-34a* regulation under serum stimulation provided proof that our assay was capable of significant sensitivity with respect to probe and target interactions.

The data normalization process determination was daunting. It was not fully understood what produced an inability for positive control sequences to reliably communicate experimental bias induced during the many stages of microRNA microarray sample preparation and labeling. Many examples of differential seeding of positive control nucleic acid were shown to reliably reproduce artificially induced normalization coefficients. These test experiments, however, were often performed independent of cellular samples and did not measure the capacity of the control vehicle to account for starting amounts or migration of endogenous microRNAs through the experimental process. In this way, positive control nucleic acids were shown to be more than capable of normalizing themselves but rarely capable of normalizing experimental samples. The strongest possible explanation for the inability of positive control normalization to directionally or numerically agree with global mean normalization is the incorrect assumption that a given mass unit of isolated total RNA has a consistent percentage of microRNA. Each microRNA microarray sample started with 100μg of total RNA subjected to size-selection for microRNA enrichment. There was an implicit assumption that within this 100μg of total RNA there was a consistent level of microRNA conserved during the process of RNA isolation. Purification columns were specifically avoided in the development of the RNA isolation procedure in an attempt to avoid any method that might intrinsically be prone to small-RNA loss. Nonetheless, differential loss of small-RNA from the total RNA population would explain both an inability for positive control normalization to operate reliably and the ability of global mean normalization to compensate.

Filtered Term Enrichment

We wished to understand what proportion of c-Myc's plethora of characterized biological roles is under the regulatory influence of c-Myc-regulated microRNAs. In order to answer this question we analyzed each binding-verified regulated microRNA for

GO term enrichment within the pool of its predicted gene targets that overlap with known c-Myc targets. In this manner we hoped to detect microRNAs that regulate gene targets that have concise biological roles that are a subcomponent of c-Myc's regulatory influence on cellular proliferation and metabolism. In the analysis of GO term enrichments for microRNA gene target sets that overlap with known c-Myc targets we intentionally ignored a large set of enrichments for ontology descriptions similar to *cell cycle* or *cellular proliferation*. These enrichments were predicted and found to be quite common with respect to the analysis of any subset of c-Myc regulated genes.

Most microRNAs showed significant enrichment of some kind. The microRNA *hsa-mir-30e*, for example, was significantly enriched for *dna repair*, *dna damage response*, and *DNA metabolism*. Additionally, there are many enriched annotations for stress response components such as *heat shock proteins*, *molecular chaperones*, and *response to unfolded proteins*. These enrichments comprise the vast bulk of all significant enrichments for the predicted gene targets of *hsa-mir-30e* regulated by c-Myc transcriptional abundance. Many publications have implicated c-Myc in the repression of DNA repair processes and cellular components [95-101]. In this manner *hsa-mir-30e* appears to provide a concise regulatory influence as a regulatory intermediary of c-Myc. The microRNA *hsa-mir-152* had singular significant enrichment for annotations corresponding to *apoptosis*, *cell death*, *death*, *regulation of programmed cell death*, and *induction of apoptosis by extracellular signals*. These annotations are related to cellular proliferation but were considered significant in this case because of the fact that no other *cell cycle* or related proliferation annotations neared statistical significance.

We believe that this is another area of analysis in which it is important to note that prediction must be filtered by cellular reality. If it were possible to take microRNAs and their pool of predicted gene targets and produce convincing ontological enrichments the research community would be much further in the process of understanding the regulatory function and biological roles of human microRNAs. The limited results we obtained from the GO term analysis performed for both serum stimulation and c-Myc modulation experiments were only possible because we were able to first experimentally produce datasets capable of filtering prediction by measurement.

Sequence Alignment

The sets of high-confidence regulated microRNAs produced for both the serum stimulation and c-Myc experiment sets were evaluated for mature microRNA sequence similarity. In addition to multiple-alignment of mature sequences, seed regions of microRNAs were separately aligned. The seed region of microRNAs is described as the consecutive stretch of 7nt starting from the first or second nucleotide of the 5' end [102]. The seed region is considered to be a major component of the regulatory interaction between a mature microRNA and the sequence with which it interacts in the 3' UTR of a gene target.

Given the target-predictive capacity of a microRNA seed region, we were interested in whether our pools of regulated microRNAs contained regulatory redundancy in the form of multiple microRNAs with matching or extremely similar seed regions. Many examples of microRNA sequence alignment were intentionally ignored in this analysis. For example, it is more than expected that several variants of a family of microRNAs would co-align (*hsa-let-7c*, *hsa-let-7e*). Several compelling examples included *hsa-mir99a/b* with *hsa-mir-100*, *hsa-mir-23a/b* with *hsa-mir-130a*, *hsa-mir-182* with *hsa-mir-96*, and *hsa-mir-148b* and *hsa-mir-152*.

The microRNAs *hsa-mir-148b* and *hsa-mir-152* were significant targets in both sets of experimentation and have perfectly matching seed regions. Of interest is the fact that c-Myc overexpression repressed *hsa-mir-148b* while *hsa-mir-152* was activated. In this manner the transcriptional abundance of microRNAs with significant target overlap are being simultaneously activated and repressed. Serum stimulation resulted in the activation of both of these microRNAs. It is possible that microRNAs are both redundant and additive with respect to their genomic dispersal and transcription response. The post-transcriptional effect of microRNA abundance may not be the pure result of activating or repressing a single microRNA. Alternatively, a unit of regulatory effect may be expressed by a specific microRNA (*hsa-mir-152*) while two units comes not from an increase in its transcriptional signal but rather the addition of transcriptional activation of a separate microRNA (*hsa-mir-148b*) with the same set of gene targets.

MATERIALS AND METHODS

Cell Culture & Experiments

Normal Cell Culture Conditions

HeLa and 2091 fibroblast cells were purchased from ATCC (American Type Culture Collection). Cultures were grown at 37°C in DMEM (Dulbecco's Modified Eagle's Medium) supplemented with 10% FBS (Fetal Bovine Serum) and 100 units penicillin-streptomycin.

2091 Fibroblast Serum Stimulation

2091 fibroblast cell cultures were grown under normal cell culture conditions until 40% confluent. Medium was removed and cell cultures were washed 3x with PBS (Phosphate Buffered Saline). Replacement medium was DMEM supplemented with 0.1% FBS and 100 units penicillin-streptomycin. Cell cultures were grown at 37°C for 48 hours. Cell cultures were washed 1x with PBS. Reference cell cultures were harvested following *Total RNA Isolation.* Replacement medium was DMEM supplemented with 10% FBS and 100 units penicillin-streptomycin. Separate cell cultures were allowed to proliferate under serum-rich conditions for time-points of 5, 10, 20, 30, 60, and 180 minutes. At the end of each of these time points cell cultures were harvested following *Total RNA Isolation.*

c-Myc Overexpression Lipotransfection

c-Myc overexpression plasmid was purchased from Open Biosystems.

Plasmid information: MHS1010-57504, Human MGC Verified FL cDNA (IRAT). CloneID=298544,ImageID=2985844,Accession=AW675223.1,Library=NIH_MGC_12, Vector=pCMV-SPORT6,Host=DH10B.

GFP co-transfection plasmid was purchased from Clontech.

Plasmid information: Vector=pEGFP-N1, Accession=U55762.

HeLa cell cultures were grown under normal cell culture conditions. 6-well plates were seeded with 1.5 x 105 cells / well. Cell cultures were allowed to grow for 24 hours. Cell cultures were transiently lipotransfected with Invitrogen Lipofectamine 2000 according to the manufacturer's protocol (for DNA plasmid transfection). Cell cultures were grown under normal cell culture conditions for 48 hours and then harvested following *Total RNA Isolation.*

c-Myc siRNA Lipotransfection

c-Myc-specific siRNA was purchased from Dharmacon.

siRNA Information: siGENOME SMARTpool deluxe (14), D-003282-14, MYC

Sense Sequence = AACGUUAGCUUCACCAACAUU

Antisense Sequence = 5’- P UGUUGGUGAAGCUAACGUUUU

siRNA Information: siGENOME SMARTpool deluxe (15), D-003282-15, MYC

Sense Sequence = GGAACUAUGACCUCGACUAUU

Antisense Sequence = 5’- P UAGUCGAGGUCAUAGUUCCUU

siRNA Information: siGENOME SMARTpool deluxe (16), D-003282-16, MYC

Sense Sequence = GAACACACAACGUCUUGGAUU

Antisense Sequence = 5’- P UCCAAGACGUUGUGUGUUCUU

siRNA Information: siGENOME SMARTpool deluxe (17), D-003282-17, MYC

Sense Sequence = GGACUAUCCUGCUGCCAAGUU

Antisense Sequence = 5’- P CUUGGCAGCAGGAUAGUCCUU

Negative control siRNA was purchased from Dharmacon.

siRNA Information: siCONTROL Non-Targeting siRNA Pool

5’ – AUGAACGUGAAUUGCUCAA – 3’

5’ – UAAGGCUAUGAAGAGAUAC – 3’

5’ – AUGUAUUGGCCUGUAUUAG – 3’

5’ – UAGCGACUAAACACAUCAA – 3’

HeLa cell cultures were grown under normal cell culture conditions. 6-well plates were seeded with 1.5 x 10^5 cells / well. Cell cultures were allowed to grow for 24 hours. Cell cultures were transiently lipotransfected with Invitrogen Lipofectamine 2000 according to the manufacturer’s protocol (for siRNA transfection). Cell cultures were grown under normal cell culture conditions for 48 hours and then harvested following *Total RNA Isolation*.

General DNA Microarray Methods

Total RNA Isolation

Invitrogen Trizol Reagent was used according to the manufacturer's protocol.

mRNA Expression Microarray Methods

Reverse Transcription

Reverse transcription of was performed using a modified Invitrogen Superscript II protocol (anchored oligo dT 5μg, total RNA 10μg). Amino Allyl dUTP was incorporated into the reverse transcription reaction for the purpose of Amersham Biosciences Cy Dye incorporation. cDNA was purified using Qiagen MinElute columns according to the manufacturer's protocol.

cDNA Fluorophore Labeling

Cy Dye incorporation was performed in the presence of purified cDNA with incorporated Amino Allyl dUTP. Cy Dyes (Cy5, Cy3) were suspended in DMSO and incubated for 60 minutes. Labeled cDNA was separated from unincorporated Cy Dyes through purification with Qiagen MinElute columns according to the manufacturer's protocol.

Microarray Slide Preparation

DNA Microarray slides were post-processed through rapid immersion and removal from 0.2 % SDS followed by 55°C incubation in 1%BSA, 5x SSC, 0.1% SDS for 45 minutes. Slides were rinsed 5x with MilliQ water and then 1x with isopropanol. Slides were spun dry in a tabletop centrifuge (600 rpm, 3 minutes).

Hybridization & Washing

Hybridization buffer consisted of 50% formamide, 10x SSC, 0.2% SDS.

Purified and labeled cDNA was combined with 2x hybridization buffer and incubated at 42°C for 16 hours. After hybridization slides were washed for 5 minutes in three stages. Stage 1 was composed of 2x SSC, 0.1%SDS (5 minutes). Stage 2 was composed of 1x SSC (5 minutes). Stage 3 was composed of 0.1x SSC (5 minutes). Slides were spun dry in a tabletop centrifuge (600 rpm, 3 minutes).

MicroRNA Expression Microarray Methods

Synthesis of Positive Control RNA

A total of 13 positive control sequences were designed and obtained from IDT.

T7 TAATACGACTCACTATAGGGAGA
C-SC-1 ATTATGCTGAGTGATATCCCTCTCTAACGGGTTTTGCGTGAACT
C-SC-1-RC ATTATGCTGAGTGATATCCCTCTAGTTCACGCAAAACCCGTTAG
C-SC-2 ATTATGCTGAGTGATATCCCTCTCCGCAGAATGGGTAAAGCTCT
C-SC-2-RC ATTATGCTGAGTGATATCCCTCTAGAGCTTTACCCATTCTGCGG
C-SC-3 ATTATGCTGAGTGATATCCCTCTTCTACAGAACACCATACTTTA
C-SC-3-RC ATTATGCTGAGTGATATCCCTCTTAAAGTATGGTGTTCTGTAGA
C-SC-4 ATTATGCTGAGTGATATCCCTCTCTGAATTAAACCTTTTGGGTT
C-SC-4-RC ATTATGCTGAGTGATATCCCTCTAACCCAAAAGGTTTAATTCAG
MIR16 ATTATGCTGAGTGATATCCCTCTATCGTCGTGCATTTATAACCGC
MIR16-RC ATTATGCTGAGTGATATCCCTCTGCGGTTATAAATGCACGACGAT
MYC-1-SENSE
ATTATGCTGAGTGATATCCCTCTGTAGTAGTAGGTCCTGACAAA

MYC-1-ANTISENSE

ATTATGCTGAGTGATATCCCTCTTGTCAGGACCTACTACTACAA

Oligos were combined with generic "T7" sequence to provide a double-stranded T7 promoter. Ambion MEGAscript T7 was used according to the manufacturer's protocol to produce single-stranded positive control RNA.

MicroRNA Enrichment, RNA Size Fractionation

Ambion FlashPAGE Fractionator System, Pre-cast Gels, Buffer Kit, and Clean-Up Kit were used according to the manufacturer's protocol.

Fluorophore Labeling, Poly-A Method

Ambion mirVana miRNA Labeling Kit was used according to the manufacturer's protocol.

Fluorophore Labeling, ULS Method

Invitrogen ULYSIS 546 Nucleic Acid Labeling Kit was used for the typical Cy3 sample. Invitrogen ULYSIS 647 Nucleic Acid Labeling Kit was used for the typical Cy5 sample. The kits were used according to the manufacturer's protocol. It was determined that ½ the recommended dye component could be used with no degradation in signal. Ambion FlashPAGE Clean-Up Kits were used for removal of unincorporated dye according to the manufacturer's protocol.

Microarray Slide Preparation

MicroRNA oligonucleotide microarrays were post-processed using a pre-prepared solution of 50ml 1M Tris, pH 9.0, 500μl 10% SDS, 300μl ethanolamine (for final concentration of 100mM). Solution was heated to 50°C. MicroRNA oligonucleotide microarrays were added to solution, sealed. Solution and slides were incubated for 20 minutes, agitating every 5 minutes. Slides were washed with MilliQ water for 1 minute. Slides were spun dry in a tabletop centrifuge (600 rpm, 3 minutes).

Hybridization & Washing

MicroRNA oligonucleotide microarrays were hybridized by combining 10μl purified and ULS-labeled microRNA samples with 20μl Ambion 3x Buffer. Microarrays were incubated at 42°C for 16 hours. After hybridization slides were washed for 1 minute in three stages. Stage 1 was composed of 2x SSC, 0.1%SDS (1 minute). Stage 2 was composed of 1x SSC (1 minute). Stage 3 was composed of 0.1x SSC (1 minute). Slides were spun dry in a tabletop centrifuge (600 rpm, 3 minutes).

Chapter 6: Conclusions and Recommendations

BIOINFORMATIC INFRASTRUCTURE

DNA microarrays have transitioned from a novel research commodity to a widespread assay capable of screening for a variety of molecular phenomena. Whole-genome expression profiling began with microarrays spotted with large PCR products capable of measuring the two-channel relative abundance of mRNAs homologous to cloned ESTSs, ORFs, and known genes. Production of microarrays now commonly involves the spotting of DNA oligonucleotides or methods such as chemical and ultra-violet lithography. Each of these technologies has increased the resolution of the assay and introduced new experimental capacity. DNA microarray experiments now include the profiling of high-resolution DNA-protein interactions, exon-specific splice-form variants, DNA copy number abnormalities, single nucleotide polymorphisms (SNPs), and regulatory chromatin modification events such as histone methylation, acetylation, phosphorylation and ubiquitination [103-108]. While the capacity of the assay has increased, the process by which biological meaning is elucidated from raw spot measurements has remained difficult and error-prone.

This difficulty and relatively high error rate is a serious problem. The average biologist spends such a significant amount of their time handling the manual aggregation of genome resources, curated gene annotations, and analytical toolsets that they easily forget the business that they are truly in; the study and understanding of biology. This problem is only accelerating as high-throughput methodologies begin to scale to even larger resulting datasets. The emerging dominance of affordable high-throughput sequencing and high-resolution whole-genome tiling microarrays will only increase the relative noise produced by experiments executed on these platforms. Finding the signal within that noise will become more difficult with this increase in scale.

The analytical insights that were achieved in Chapter 4 (*A Functional Transcriptional Regulatory Network*) are representative of what is possible when more time is spent defining and answering biological questions than struggling with technology. For each result presented, there were at least three other large-scale questions that were asked, answered, and then set aside. This was possible because we had developed both a robust data model and a set of genome resources and analytical toolsets that could rapidly interact with it. The consistent use of directed weighted graphs, the use of a repeatable pipeline of analytical procedures, and the availability of a programmatic API by which updated genome resources could be interacted with greatly shortened the cycle of think, ask, test, and answer. That is truly the business of the biologist – think, ask, test, answer. Think, ask, test, and answer.

The toolsets and technologies presented in Chapter 3, however, represent only a temporary reprieve for the experimental biologists at work in the field of functional genomics. The Longhorn Array Database and ArrayPlex are tools of the status quo. They have great capacity to handle the common bioinformatic problems of today. They too, however, are unprepared for tomorrow. Computational biologists need to begin the process of leading the reorganization and standardization of the vast data resources and toolsets that will be central to the next generation of experimental methods, analytical questions, and resulting sets of answers. Standards such as MIAME and MAGE should not be proposed, published, supported, and then somewhat forgotten when they become uninteresting or seem untenable. Experimental biologists would certainly be the beneficiaries but will not be the champions of these technological causes. Their loyalties, understandably, are to results, grants, thinking, asking, testing, and answering. This is as it should be. This is the business of experimental biology. The business, however, of bioinformatics and computational biology is one of both providing analytical empowerment now and preparing for that capacity in the future. The coming emergence of the next generation of high-throughput experimental technologies and results demands more attention is paid to the universal portability of gene and protein identifiers, genome sequence revisions, hierarchical and organism-independent ontology identifiers, and the ability for analytical toolsets to portably share input and output analytical connections in more productive, functional ways.

MICRORNA TRANSCRIPTIONAL ABUNDANCE

The computational prediction of microRNA targets is necessary to even begin the process of mechanistic *in vivo* validation. Most human microRNAs, however, have more than 1000 predicted gene targets [62, 69, 109]. These regulatory predictions must be functionally supplemented with experimental characterization in order to have any hope of obtaining regulatory signal from the omnipresent noise. The regulatory analysis and predictions performed for Chapter 5 of this study were sensitive to the directionality of microRNAs and gene targets because experimental design allowed us to characterize the cellular behavior of both regulator and target under matching physiological conditions. In this manner we were able to take the relatively weak regulatory signal of putative prediction and strengthen it through measured activation of microRNA and concomitant repression of target. Similarly, we were able to take the unconvincing presence of upstream transcription factor motifs and strengthen it through measured activation of transcription factor and microRNA transcriptional response.

Both the serum stimulation and c-Myc modulation experiments demonstrated capacity for regulatory motif depletion in promoter regions. The serum stimulation experiments showed that sequence motifs for transcription factors activated by serum stimulation were significantly depleted from promoter regions of microRNAs that were repressed under the same conditions. In a similar manner, the non-canonical E-box variant (CATGTG) was 75% depleted in the set of microRNAs activated by serum stimulation. The c-Myc experiments demonstrate very similar results with respect to the depletion of this non-canonical variant. MicroRNAs activated by c-Myc overexpression were specifically depleted for this motif while repressed microRNAs showed significant over-occurrence. It is known that sequence motif gain and loss is a major component of the evolutionary elasticity of microRNA-mediated regulation on gene targets [110]. The results here show that similar evolutionary gain and loss may be in effect to either accentuate or prevent the transcriptional activation of microRNAs under certain cellular conditions.

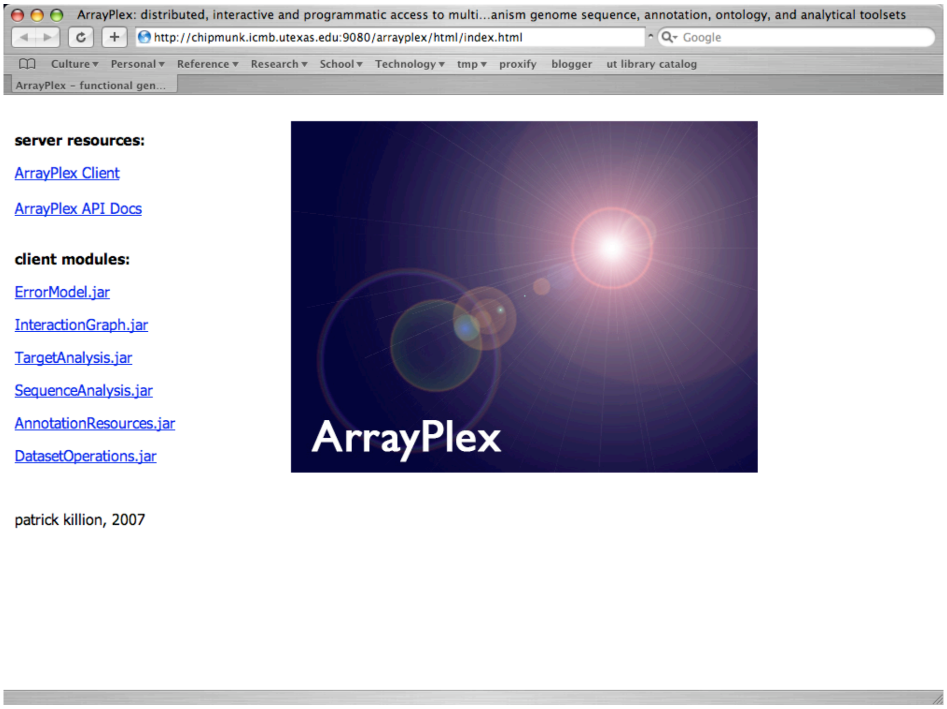
Further research is needed to accelerate the characterization of microRNA transcriptional regulation. The elucidation of transcription factor mediated regulation on gene targets has been a challenging process that will continue to provide many research opportunities for decades to come. The continued emergence of new principles such as the histone code will forever keep complacency in check with respect to the battle of high-throughput molecular biology experiments vs. the outstanding complexity of mammalian transcriptional regulation [111]. More understanding is needed of the basic nature of microRNA primary transcript maturation. Very few microRNAs have mapped transcriptional start sites. Polycistronic clusters of microRNAs are often predicted but still not characterized. The very definition of microRNA core promoters and enhancers is still at its relative infancy when compared to the relative information compiled for protein-coding gene transcription.

Finally, the biogenesis of microRNAs is not simply confined to transcriptional activation and repression. Recent studies have demonstrated that non-transcriptional nuclear and cytoplasmic factors may play significant roles in the sequestration or availability of mature microRNA sequences [68, 112-114]. The results of our serum stimulation experiments were interesting in this regard. A significant subset of the microRNAs regulated by serum stimulation was also regulated by c-Myc modulation. This was not a surprise as c-Myc is activated by serum stimulation. What was surprising and not fully understood was the temporal speed by which many of the microRNAs were both activated and repressed. The first time-point in the serum stimulation experiments was 5 minutes. Many microRNAs were regulated within this period and one was validated with qPCR. This time frame, however, likely precedes the transcriptional activation of c-Myc itself. The biogenesis mechanisms that allow this rapid maturation of microRNA need further investigation in order to truly begin to understand the nature of microRNA regulation.

Appendix I – ArrayPlex Client Feature Set

The ArrayPlex Client operates on Apple Mac OS X, Microsoft Windows (XP/Vista), and all distributions of the Linux operating systems.

AI-1 – ArrayPlex Server Launch Screen


ArrayPlex: distributed, interactive and programmatic access to multi...anism genome sequence, annotation, ontology, and analytical toolsets
http://chipmunk.icmb.utexas.edu:9080/arrayplex/html/index.html
Google
Culture
Personal
Reference
Research
School
Technology
tmp
proxify
blogger
ut library catalog
ArrayPlex – functional gen...
ArrayPlex

server resources:

ArrayPlex Client

ArrayPlex API Docs

client modules:

ErrorModel.jar

InteractionGraph.jar

TargetAnalysis.jar

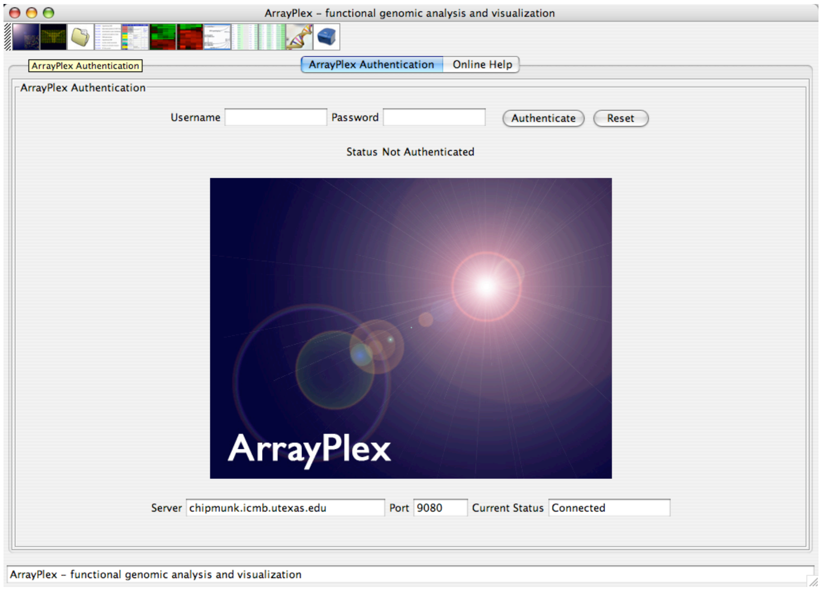
SequenceAnalysis.jar

AnnotationResources.jar

DatasetOperations.jar

patrick killion, 2007

AI-2 – ArrayPlex Client Authentication



Status Not Authenticated

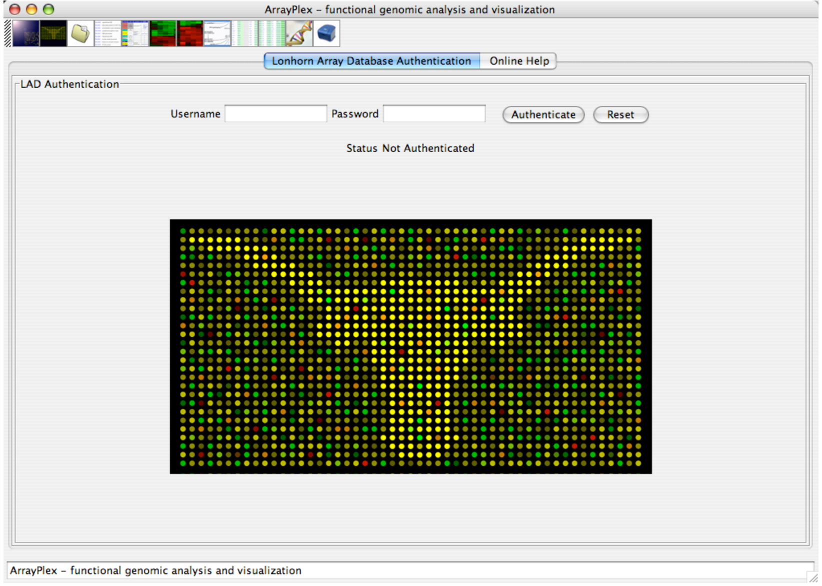
Server chipmunk.icmb.utexas.edu

Port 9080

Current Status Connected

ArrayPlex – functional genomic analysis and visualization

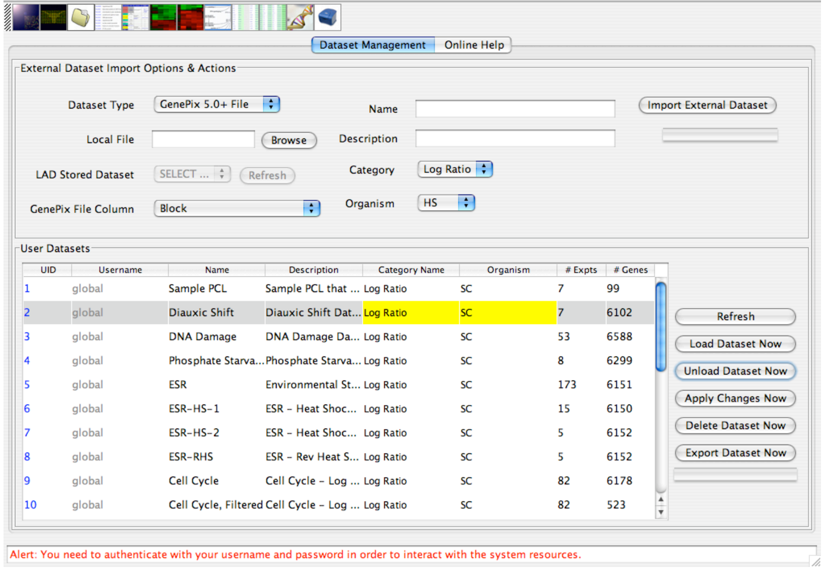
AI-3 – ArrayPlex Client Longhorn Array Database Authentication


ArrayPlex – functional genomic analysis and visualization
Lonhorn Array Database Authentication
Online Help
LAD Authentication
Username
Password
Authenticate
Reset

Status Not Authenticated

ArrayPlex – functional genomic analysis and visualization

AI-4 – ArrayPlex Client Dataset Management

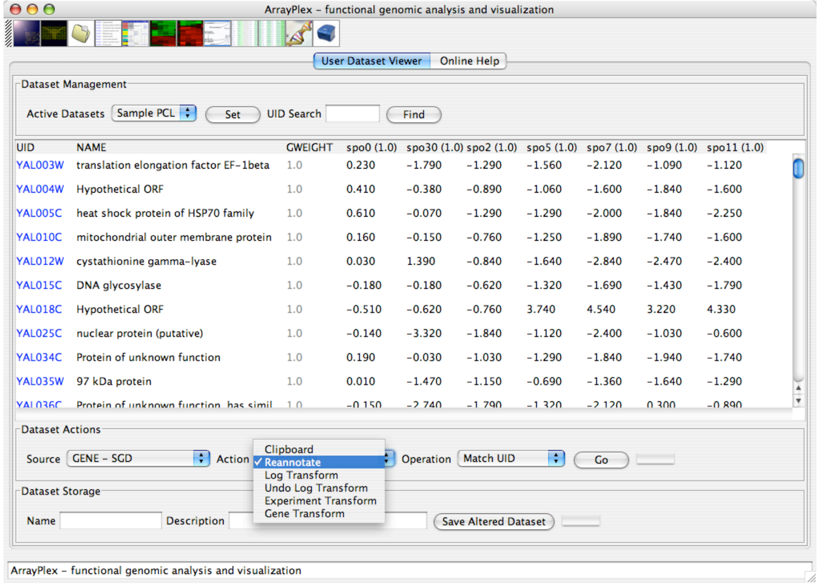


Dataset Type: GenePix 5.0+ File
Local File:
LAD Stored Dataset: SELECT ...
GenePix File Column: Block
Name:
Description:
Category: Log Ratio
Organism: HS

UID	Username	Name	Description	Category Name	Organism	# Expts	# Genes
1	global	Sample PCL	Sample PCL that ...	Log Ratio	SC	7	99
2	global	Diauxic Shift	Diauxic Shift Dat...	Log Ratio	SC	7	6102
3	global	DNA Damage	DNA Damage Da...	Log Ratio	SC	53	6588
4	global	Phosphate Starva...	Phosphate Starva...	Log Ratio	SC	8	6299
5	global	ESR	Environmental St...	Log Ratio	SC	173	6151
6	global	ESR-HS-1	ESR – Heat Shoc...	Log Ratio	SC	15	6150
7	global	ESR-HS-2	ESR – Heat Shoc...	Log Ratio	SC	5	6152
8	global	ESR-RHS	ESR – Rev Heat S...	Log Ratio	SC	5	6152
9	global	Cell Cycle	Cell Cycle – Log ...	Log Ratio	SC	82	6178
10	global	Cell Cycle, Filtered	Cell Cycle – Log ...	Log Ratio	SC	82	523

Alert: You need to authenticate with your username and password in order to interact with the system resources.

AI-5 – ArrayPlex Client Dataset Display



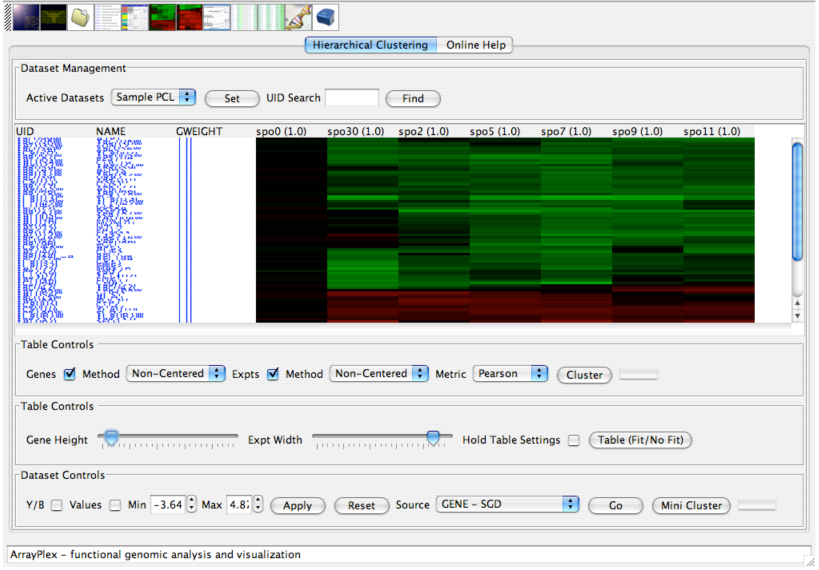
Active Datasets: Sample PCL
UID Search:

UID	NAME	GWEIGHT	spo0 (1.0)	spo30 (1.0)	spo2 (1.0)	spo5 (1.0)	spo7 (1.0)	spo9 (1.0)	spo11 (1.0)
YAL003W	translation elongation factor EF-1beta	1.0	0.230	-1.790	-1.290	-1.560	-2.120	-1.090	-1.120
YAL004W	Hypothetical ORF	1.0	0.410	-0.380	-0.890	-1.060	-1.600	-1.840	-1.600
YAL005C	heat shock protein of HSP70 family	1.0	0.610	-0.070	-1.290	-1.290	-2.000	-1.840	-2.250
YAL010C	mitochondrial outer membrane protein	1.0	0.160	-0.150	-0.760	-1.250	-1.890	-1.740	-1.600
YAL012W	cystathionine gamma-lyase	1.0	0.030	1.390	-0.840	-1.640	-2.840	-2.470	-2.400
YAL015C	DNA glycosylase	1.0	-0.180	-0.180	-0.620	-1.320	-1.690	-1.430	-1.790
YAL018C	Hypothetical ORF	1.0	-0.510	-0.620	-0.760	3.740	4.540	3.220	4.330
YAL025C	nuclear protein (putative)	1.0	-0.140	-3.320	-1.840	-1.120	-2.400	-1.030	-0.600
YAL034C	Protein of unknown function	1.0	0.190	-0.030	-1.030	-1.290	-1.840	-1.940	-1.740
YAL035W	97 kDa protein	1.0	0.010	-1.470	-1.150	-0.690	-1.360	-1.640	-1.290
YAL036C	Protein of unknown function, has simil	1.0	-0.150	-2.740	-1.790	-1.320	-2.120	0.300	-0.890

Source: GENE – SGD
Action: Reannotate
Operation: Match UID
Name:
Description:

ArrayPlex – functional genomic analysis and visualization

AI-6 – ArrayPlex Client Hierarchical Clustering



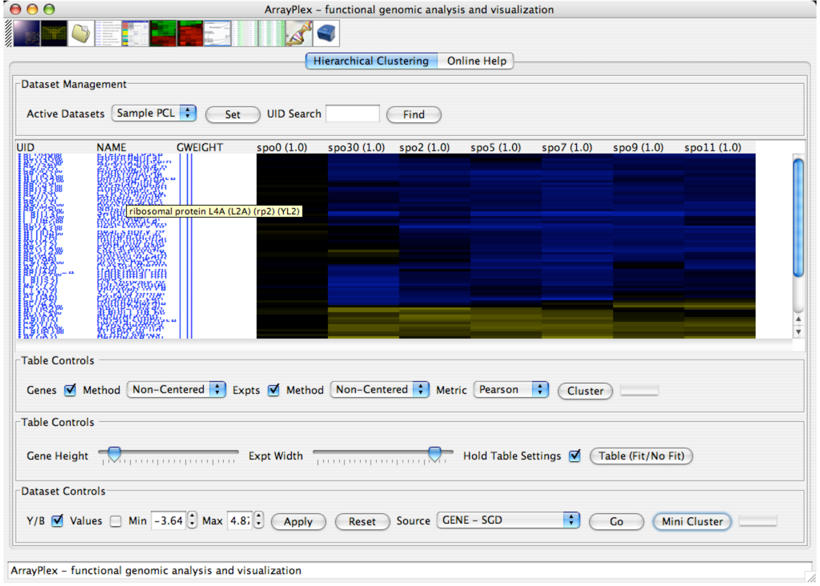
Active Datasets: Sample PCL

Genes ☑ Method: Non-Centered; Expts ☑ Method: Non-Centered; Metric: Pearson

Hold Table Settings ☐

Y/B ☐ Values ☐ Min: -3.64 Max: 4.8; Source: GENE – SGD

AI-7 – ArrayPlex Client Hierarchical Clustering (RG Color-Blind Mode)



ribosomal protein L4A (L2A) (rp2) (YL2)

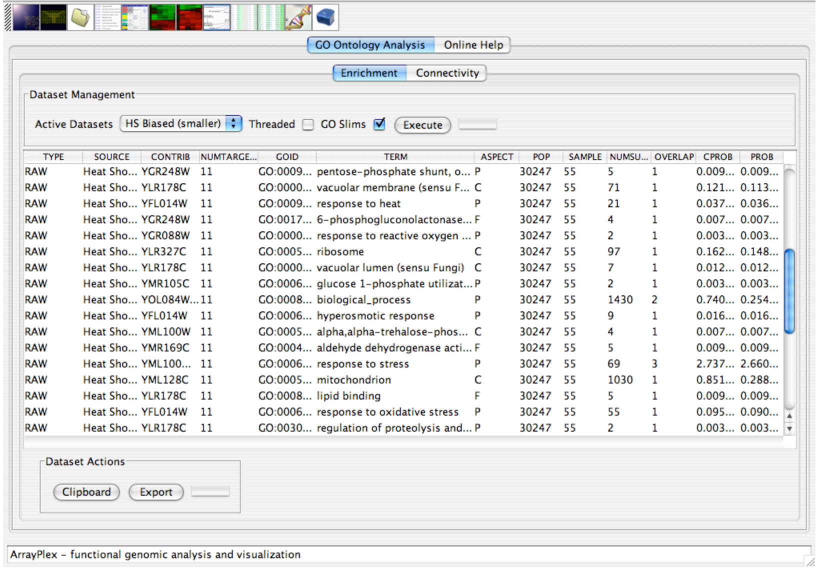
Active Datasets: Sample PCL

Genes ☑ Method: Non-Centered; Expts ☑ Method: Non-Centered; Metric: Pearson

Hold Table Settings ☑

Y/B ☑ Values ☐ Min: -3.64 Max: 4.8; Source: GENE – SGD

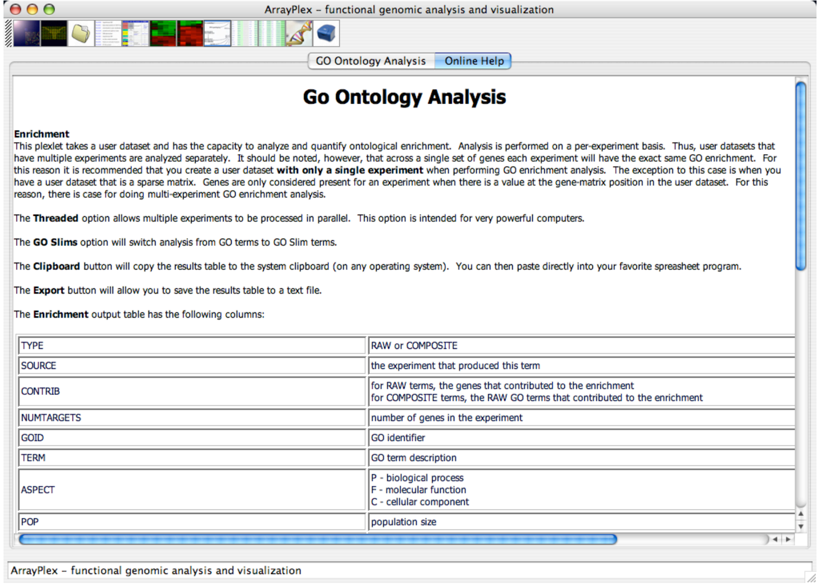
AI-8 – ArrayPlex Client GO Ontology Enrichment Analysis


ArrayPlex – functional genomic analysis and visualization
GO Ontology Analysis
Online Help
Enrichment
Connectivity
Dataset Management
Execute
Dataset Actions
Clipboard
Export
ArrayPlex – functional genomic analysis and visualization

Active Datasets: HS Biased (smaller)   Threaded ☐   GO Slims ☑

TYPE	SOURCE	CONTRIB	NUMTARGE...	GOID	TERM	ASPECT	POP	SAMPLE	NUMSU...	OVERLAP	CPROB	PROB
RAW	Heat Sho...	YGR248W	11	GO:0009...	pentose-phosphate shunt, o...	P	30247	55	5	1	0.009...	0.009...
RAW	Heat Sho...	YLR178C	11	GO:0000...	vacuolar membrane (sensu F...	C	30247	55	71	1	0.121...	0.113...
RAW	Heat Sho...	YFL014W	11	GO:0009...	response to heat	P	30247	55	21	1	0.037...	0.036...
RAW	Heat Sho...	YGR248W	11	GO:0017...	6-phosphogluconolactonase...	F	30247	55	4	1	0.007...	0.007...
RAW	Heat Sho...	YGR088W	11	GO:0000...	response to reactive oxygen ...	P	30247	55	2	1	0.003...	0.003...
RAW	Heat Sho...	YLR327C	11	GO:0005...	ribosome	C	30247	55	97	1	0.162...	0.148...
RAW	Heat Sho...	YLR178C	11	GO:0000...	vacuolar lumen (sensu Fungi)	C	30247	55	7	1	0.012...	0.012...
RAW	Heat Sho...	YMR105C	11	GO:0006...	glucose 1-phosphate utilizat...	P	30247	55	2	1	0.003...	0.003...
RAW	Heat Sho...	YOL084W...	11	GO:0008...	biological_process	P	30247	55	1430	2	0.740...	0.254...
RAW	Heat Sho...	YFL014W	11	GO:0006...	hyperosmotic response	P	30247	55	9	1	0.016...	0.016...
RAW	Heat Sho...	YML100W	11	GO:0005...	alpha,alpha-trehalose-phos...	C	30247	55	4	1	0.007...	0.007...
RAW	Heat Sho...	YMR169C	11	GO:0004...	aldehyde dehydrogenase acti...	F	30247	55	5	1	0.009...	0.009...
RAW	Heat Sho...	YML100...	11	GO:0006...	response to stress	P	30247	55	69	3	2.737...	2.660...
RAW	Heat Sho...	YML128C	11	GO:0005...	mitochondrion	C	30247	55	1030	1	0.851...	0.288...
RAW	Heat Sho...	YLR178C	11	GO:0008...	lipid binding	F	30247	55	5	1	0.009...	0.009...
RAW	Heat Sho...	YFL014W	11	GO:0006...	response to oxidative stress	P	30247	55	55	1	0.095...	0.090...
RAW	Heat Sho...	YLR178C	11	GO:0030...	regulation of proteolysis and...	P	30247	55	2	1	0.003...	0.003...

AI-9 – ArrayPlex Client GO Ontology Enrichment Documentation


ArrayPlex – functional genomic analysis and visualization
GO Ontology Analysis
Online Help
ArrayPlex – functional genomic analysis and visualization

Go Ontology Analysis

Enrichment
This plexlet takes a user dataset and has the capacity to analyze and quantify ontological enrichment. Analysis is performed on a per-experiment basis. Thus, user datasets that have multiple experiments are analyzed separately. It should be noted, however, that across a single set of genes each experiment will have the exact same GO enrichment. For this reason it is recommended that you create a user dataset **with only a single experiment** when performing GO enrichment analysis. The exception to this case is when you have a user dataset that is a sparse matrix. Genes are only considered present for an experiment when there is a value at the gene-matrix position in the user dataset. For this reason, there is case for doing multi-experiment GO enrichment analysis.

The **Threaded** option allows multiple experiments to be processed in parallel. This option is intended for very powerful computers.

The **GO Slims** option will switch analysis from GO terms to GO Slim terms.

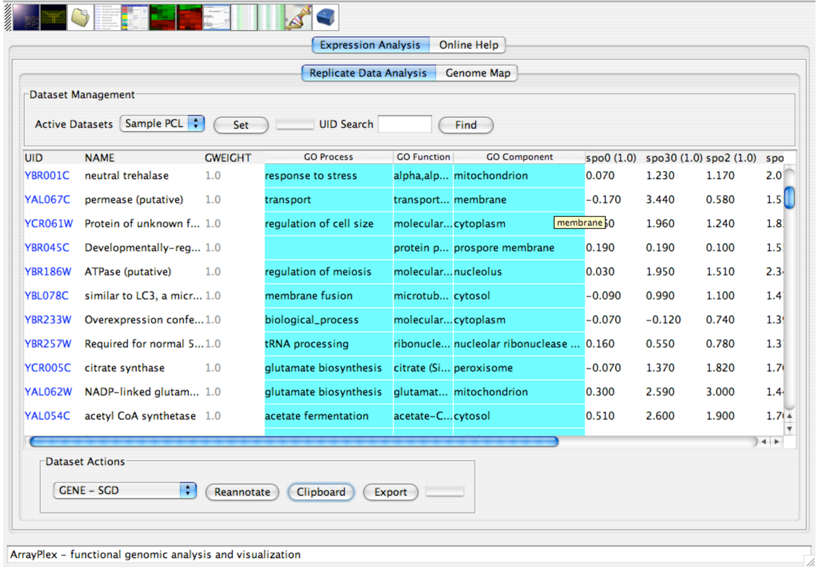
The **Clipboard** button will copy the results table to the system clipboard (on any operating system). You can then paste directly into your favorite spreasheet program.

The **Export** button will allow you to save the results table to a text file.

The **Enrichment** output table has the following columns:

TYPE	RAW or COMPOSITE
SOURCE	the experiment that produced this term
CONTRIB	for RAW terms, the genes that contributed to the enrichment for COMPOSITE terms, the RAW GO terms that contributed to the enrichment
NUMTARGETS	number of genes in the experiment
GOID	GO identifier
TERM	GO term description
ASPECT	P - biological process F - molecular function C - cellular component
POP	population size

AI-10 – ArrayPlex Client Replicate Experiment Processing



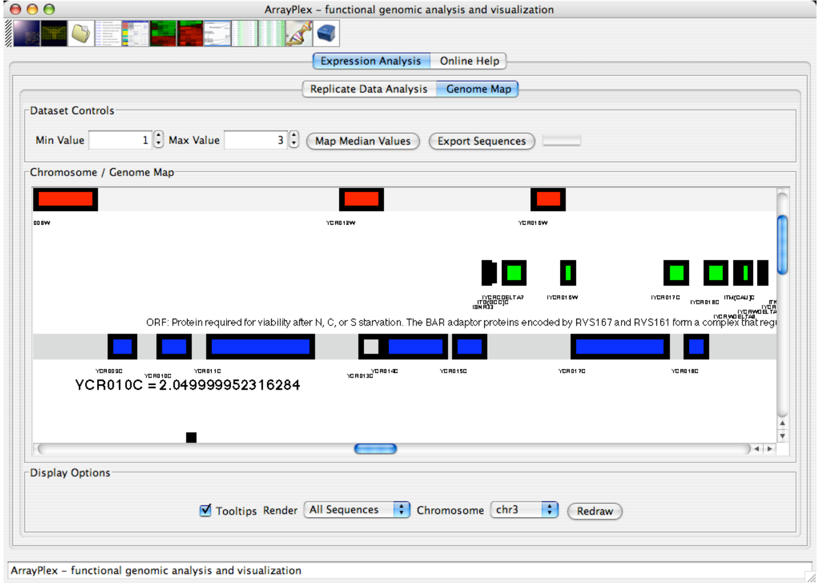
Active Datasets: Sample PCL

UID Search:

UID	NAME	GWEIGHT	GO Process	GO Function	GO Component	spo0 (1.0)	spo30 (1.0)	spo2 (1.0)	spo
YBR001C	neutral trehalase	1.0	response to stress	alpha,alp...	mitochondrion	0.070	1.230	1.170	2.0
YAL067C	permease (putative)	1.0	transport	transport...	membrane	-0.170	3.440	0.580	1.5
YCR061W	Protein of unknown f...	1.0	regulation of cell size	molecular...	cytoplasm	0	1.960	1.240	1.8
YBR045C	Developmentally-reg...	1.0		protein p...	prospore membrane	0.190	0.190	0.100	1.5
YBR186W	ATPase (putative)	1.0	regulation of meiosis	molecular...	nucleolus	0.030	1.950	1.510	2.3
YBL078C	similar to LC3, a micr...	1.0	membrane fusion	microtub...	cytosol	-0.090	0.990	1.100	1.4
YBR233W	Overexpression confe...	1.0	biological_process	molecular...	cytoplasm	-0.070	-0.120	0.740	1.3
YBR257W	Required for normal 5...	1.0	tRNA processing	ribonucle...	nucleolar ribonuclease ...	0.160	0.550	0.780	1.3
YCR005C	citrate synthase	1.0	glutamate biosynthesis	citrate (Si...	peroxisome	-0.070	1.370	1.820	1.7
YAL062W	NADP-linked glutam...	1.0	glutamate biosynthesis	glutamat...	mitochondrion	0.300	2.590	3.000	1.4
YAL054C	acetyl CoA synthetase	1.0	acetate fermentation	acetate-C...	cytosol	0.510	2.600	1.900	1.7

GENE – SGD

AI-11 – ArrayPlex Client *Saccharomyces cerevisiae* Genome Mapping



Min Value 1

Max Value 3

ORF: Protein required for viability after N, C, or S starvation. The BAR adaptor proteins encoded by RVS167 and RVS161 form a complex that regu

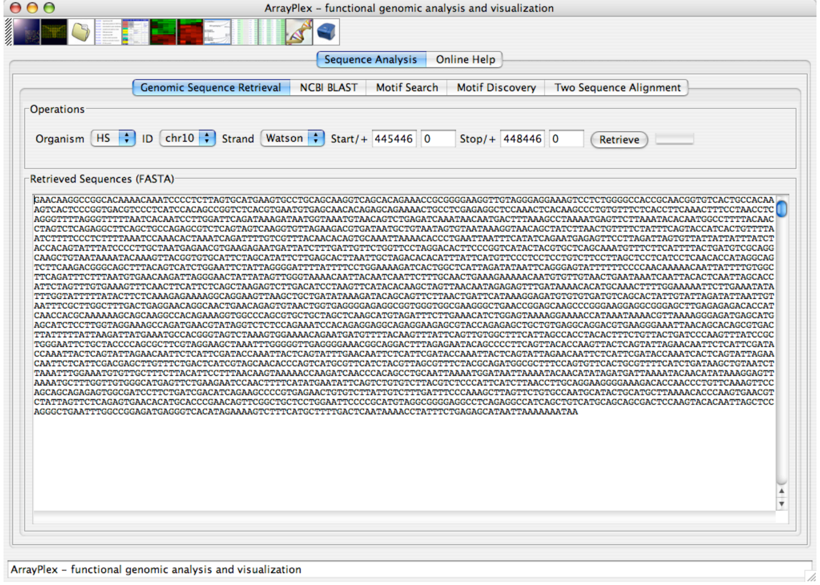
YCR010C = 2.049999952316284

☑ Tooltips

Render: All Sequences

Chromosome: chr3

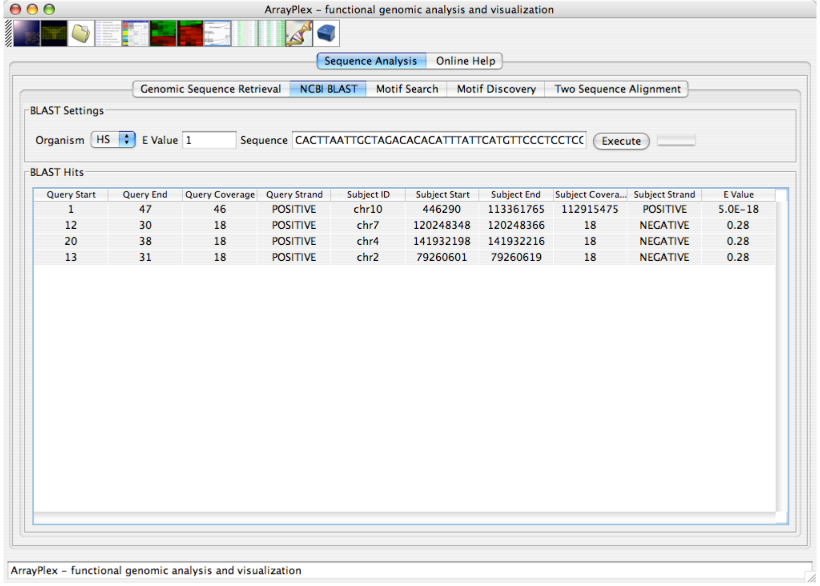
AI-12 – ArrayPlex Client Authentication Genome Sequence Retrieval


ArrayPlex – functional genomic analysis and visualization
Sequence Analysis
Online Help
Genomic Sequence Retrieval
NCBI BLAST
Motif Search
Motif Discovery
Two Sequence Alignment
Operations
Retrieve
Retrieved Sequences (FASTA)
ArrayPlex – functional genomic analysis and visualization

Organism HS ID chr10 Strand Watson Start/+ 445446 0 Stop/+ 448446 0

```
GAACAAGGCCGGCACAAAACAAATCCCCTCTTAGTGCATGAAGTGCCTGCAGCAAGGTCAGCACAGAAACCGCGGGGAAGGTTGTAGGGAGGAAAGTCCTCTGGGGCCACCGCAACGGTGTCACTGCCACAA
AGTCACTCCCGGTGACGTCCCTCATCCACAGCCGGTCTCACGTGAATGTGAGCAACACAGAGCAGAAAACTGCCTCGAGAGGCTCCAAACTCACAAGCCCTGTGTTTCTCACCTTCAAACTTTCCTAACCTC
AGGGTTTTAGGGTTTTTAATCACAATCCTTGGATTCAGATAAAGATAATGGTAAATGTAACAGTCTGAGATCAAATAACAATGACTTTAAAGCCTAAAATGAGTTCTTAAATACACAATGGCCTTTTACAAC
CTAGTCTCAGAGGCTTCAGCTGCCAGAGCGTCTCAGTAGTCAAGGTGTTAGAAGACGTGATAATGCTGTAATAGTGTAATAAAGGTAACAGCTATCTTAACTGTTTTCTATTTCAGTACCATCACTGTTTTA
ATCTTTTCCCTCTTTTAAATCCAAACACTAAATCAGATTTTGTCGTTTACAACACAGTGCAAATTAAAACACCCTGAATTAATTTCATATCAGAATGAGAGTTCCTTAGATTAGTGTTATTATTATTTATCT
ACCACAGTATTTATCCCCTTGCTAATGAGAACGTGAAGAGAATGATTATCTTTGATTGTTCTGGTTCCTAGGACACTTCCCGGTCATACTACGTGCTCAGCAAATGTTTCTTCATTTTACTGATGTCGCAGG
CAAGCTGTAATAAAATACAAAGTTACGGTGTGCATTCTAGCATATTCTTGAGCACTTAATTGCTAGACACACATTTATTCATGTTCCCTCCTCCTGTCTTCCTTAGCTCCTCATCCTCAACACCATAGGCAG
TCTTCAAGACGGGCAGCTTTACAGTCATCTGGAATTCTATTAGGGGATTTTATTTTCCTGGAAAAGATCACTGGCTCATTAGATATAATTCAGGGAGTATTTTTTCCCCAACAAAAACAATTATTTTGTGGC
TTCAGATTTCTTTAATGTGAACAAGATTAGGGAACTATTATAGTTGGGTAAAACAATTACAATCAATTCTTTGCAACTGAAAGAAAAACAATGTGTTGTAACTGAATAAATCAATTACACTCAATTAGCACC
ATTCTAGTTTGTGAAAGTTTCAACTTCATTCTCAGCTAAGAGTCTTGACATCCTAAGTTCATACACAAGCTAGTTAACAATAGAGAGTTTGATAAAACACATGCAAACTTTTGGAAAAATTCTTGAAATATA
TTTGGTATTTTTATACTTCTCAAAGAGAAAAGGCAGGAAGTTAAGCTGCTGATATAAAGATACAGCAGTTCTTAACTGATTCATAAAGGAGATGTGTGTGATGTCAGCACTATTGTATTAGATATTAATTGT
AATTTCGCTTGGCTTTGACTGAGGAACAGGCAACTGAACAGAGTGTAACTGGTGAGGGGAGAGGCGGTGGGTGGCGAAGGGCTGAACCGGAGCAAGCCCGGGAAGGAGGCGGGAGCTTGAGAGAGACACCAT
CAACCACGCAAAAAAGCAGCAAGGCCACAGAAAGGTGGCCCAGCGTGCTGCTAGCTCAAGCATGTAGATTTCTTGAAACATCTGGAGTAAAAGGAAAACCATAAATAAAACGTTAAAAGGGAGATGAGCATG
AGCATCTCCTTGGTAGGAAAGCCAGATGAACGTATAGGTCTCTCCAGAAATCCACAGAGGAGGCAGAGGAAGAGCGTACCAGAGAGCTGCTGTGAGGCAGGACGTGAAGGGAAATTAACAGCACAGCGTGAC
TTATTTTTATTAAGATTATGAAATGCCACGGGTAGTCTAAAGTGGAAAACAGAATGATGTTTTACAAGTTTATTCAGTTGTGGCTTTCATTAGCCACCTACACTTTCTGTTACTGATCCCAAGTTTATCCGC
TGGGAATTCTGCTACCCCAGCGCTTCGTAGGAAGCTAAATTTGGGGGTTGAGGGGAAACGGCAGGACTTTAGAGAATACAGCCCCTTCAGTTACACCAAGTTACTCAGTATTAGAACAATTCTCATTCGATA
CCAAATTACTCAGTATTAGAACAATTCTCATTCGATACCAAATTACTCAGTATTTGAACAATTCTCATTCGATACCAAATTACTCAGTATTAGAACAATTCTCATTCGATACCAAATCACTCAGTATTAGAA
CAATTCTCATTCGACGAGCTTGTTTCTGACTCATCGTAGCAACACCCAGTCATGCGTTCATCTACGTTAGCGTTTCTACGCAGATGGCGCTTTCCAGTGTTCACTGCGTTTTCATCTGATAAGCTGTAATCT
TAAATTTGGAAATGTGTTGCTTTCTTACATTCCTTTAACAAGTAAAAACCAAGATCAACCCACAGCCTGCAATTAAAATGGATAATTAAAATACAACATATAGATGATTAAAATACAACATATAAAGGAGTT
AAAATGCTTTGGTTGTGGGCATGAGTTCTGAAGAATCCAACTTTTCATATGAATATTCAGTCTGTGTCTTACGTCTCCCATTCATCTTAACCTTGCAGGAAGGGGAAAGACACCAACCCTGTTCAAAGTTCC
AGCAGCAGAGAGTGGCGATCCTTCTGATCGACATCAGAAGCCCCGTGAGAACTGTGTCTTATTGTCTTTGATTTCCCAAAGCTTAGTTCTGTGCCAATGCATACTGCATGCTTAAAACACCCAAGTGAACGT
CTATTAGTTCTCAGAGTGAACACATGCACCCGAACAGTTCGGCTGCTCCTGGAATTCCCCGCATGTAGGCGGGGAGGCCTCAGAGGCCATCAGCTGTCATGCAGCAGCGACTCCAAGTACACAATTAGCTCC
AGGGCTGAATTTGGCCGGAGATGAGGGTCACATAGAAAAGTCTTTCATGCTTTTGACTCAATAAAACCTATTTCTGAGAGCATAATTAAAAAAAATAA
```

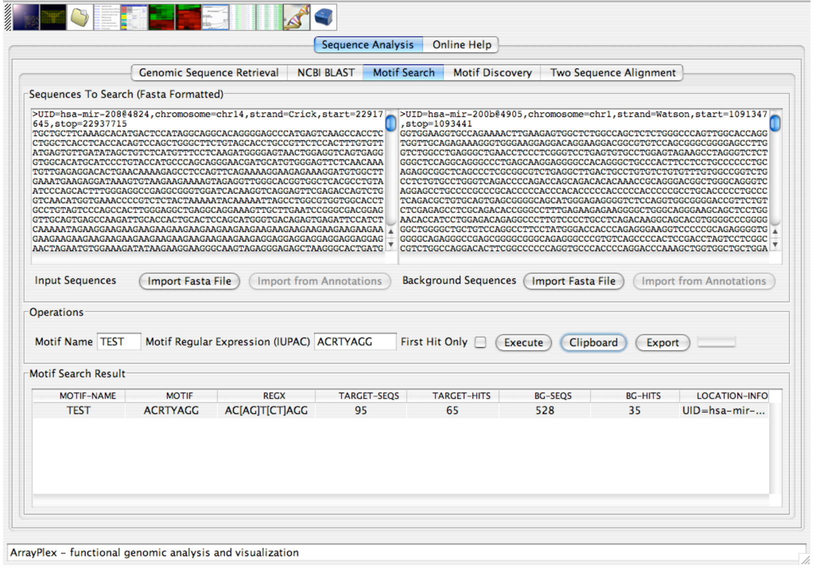
AI-13 – ArrayPlex Client BLAST


ArrayPlex – functional genomic analysis and visualization
Sequence Analysis
Online Help
Genomic Sequence Retrieval
NCBI BLAST
Motif Search
Motif Discovery
Two Sequence Alignment
BLAST Settings
Execute
BLAST Hits
ArrayPlex – functional genomic analysis and visualization

Organism HS E Value 1 Sequence CACTTAATTGCTAGACACACATTTATTCATGTTCCCTCCTCC

Query Start	Query End	Query Coverage	Query Strand	Subject ID	Subject Start	Subject End	Subject Covera...	Subject Strand	E Value
1	47	46	POSITIVE	chr10	446290	113361765	112915475	POSITIVE	5.0E-18
12	30	18	POSITIVE	chr7	120248348	120248366	18	NEGATIVE	0.28
20	38	18	POSITIVE	chr4	141932198	141932216	18	NEGATIVE	0.28
13	31	18	POSITIVE	chr2	79260601	79260619	18	NEGATIVE	0.28

AI-14 – ArrayPlex Client Authentication Sequence Motif Search



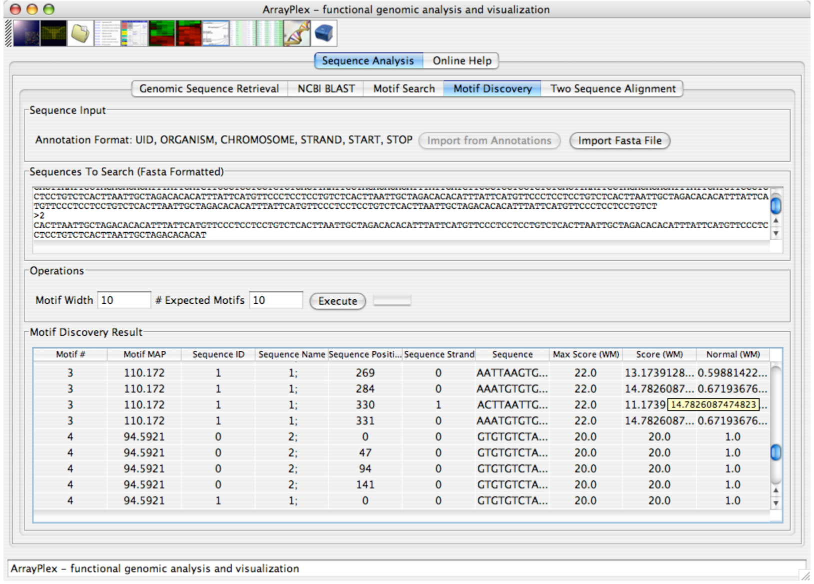
```
>UID=hsa-mir-208@4824,chromosome=chr14,strand=Crick,start=22917
645,stop=22937715
TGCTGCTTCAAAGCACATGACTCCATAGGCAGGCACAGGGGAGCCCATGAGTCAAGCCACCTC
CTGGCTCACCTCACCACAGTCCAGCTGGGCTTCTGTAGCACCTGCCGTTCTCCACTTTGTGTT
ATGAGTGTTGATATAGCTGTCTCATGTTTCCTCAAGATGGGGAGTAACTGGAGGTCAGTGAGG
GTGGCACATGCATCCCTGTACCATGCCCAGCAGGGAACGATGCATGTGGGAGTTCTCAACAAA
TGTTGAGAGGACACTGAACAAAAGAGCCTCCAGTTCAGAAAAGGAAGAGAAAGGATGTGGCTT
GAAATGAAGAGGATAAAGTGTAAGAAGAAAAGTAGAGGTTGGGCACGGTGGCTCACGCCTGTA
ATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCACAAGGTCAGGAGTTCGAGACCAGTCTG
GTCAACATGGTGAAACCCCGTCTCTACTAAAAATACAAAAATTAGCCTGGCGTGGTGGCACCT
GCCTGTAGTCCCAGCCACTTGGGAGGCTGAGGCAGGAAAGTTGCTTGAATCCGGGCGACGGAG
GTTGCAGTGAGCCAAGATTGCACCACTGCACTCCAGCATGGGTGACAGAGTGAGATTCCATCT
CAAAAATAGAAGGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAA
GAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAGGAGGAGGAGGAGGAGGAGGAG
AACTAGAATGTGGAAAGATATAAGAAGGAAGGGCAAGTAGAGGGAGAGCTAAGGGCACTGATG
```

```
>UID=hsa-mir-200b@4905,chromosome=chr1,strand=Watson,start=1091347
,stop=1093441
GGTGGAAGGTGCCAGAAAACTTGAAGAGTGGCTCTGGCCAGCTCTCTGGGCCCAGTTGGCACCAGG
TGGTTGCAGAGAAAGGGTGGGAAGGAGGACAGGAAGGACGGCGTGTCCAGCGGGCGGGGAGCCTTG
GTCTGGCCTGAGGGCTGAACCTCCCTCGGGTCCTGAGTGTGCCTGGAGTAGAAGCCTAGGGTCTCT
GGGCTCCAGGCAGGGCCCTGAGCAAGGAGGGGCCACAGGGCTGCCCACTTCCTCCTGCCCCCCTGC
AGAGGCGGCTCAGCCCTCGCGGCGTCTGAGGCTTGACTGCCTGTGTCTGTGTTTGTGGCCGGTCTG
CCTCTGTGCCTGGGTCAGACCCCAGACCAGCAGACACACAAACCGCAGGGACGGCTGGGCAGGGTC
AGGAGCCTGCCCCGCCCGCACCCCCACCCACACCCCCACCCCCACCCCCGCCTGCACCCCCTGCCC
TCAGACGCTGTGCAGTGAGCGGGGCAGCATGGGAGAGGGGTCTCCAGGTGGCGGGGACCGTTCTGT
CTCGAGAGCCTCGCAGACACCGGGCCTTTGAGAAGAGAAGGGGCTGGGCAGGAAGCAGCTCCTGG
AACACCATCCTGGAGACAGAGGCCCTTGTCCCCTGCCTCAGACAAGGCAGCACGTGGGGCCCGGGG
GGCTGGGGCTGCTGTCCAGGCCTTCCTATGGACCACCCAGAGGGAAGGTCCCCCGCAGAGGGGTG
GGGGCAGAGGGCCGAGCGGGGCGGGCAGAGGGCCCGTGTCAGCCCCACTCCGACCTAGTCCTCGGC
CGTCTGGCCAGGACACTTCGGCCCCCCAGGTGCCCACCCCAGGACCCAAAGCTGGTGGCTGCTGGA
```

Motif Name TEST Motif Regular Expression (IUPAC) ACRTYAGG First Hit Only ☐

MOTIF-NAME	MOTIF	REGX	TARGET-SEQS	TARGET-HITS	BG-SEQS	BG-HITS	LOCATION-INFO
TEST	ACRTYAGG	AC[AG]T[CT]AGG	95	65	528	35	UID=hsa-mir-...

AI-15 – ArrayPlex Client Authentication Sequence Motif Discovery



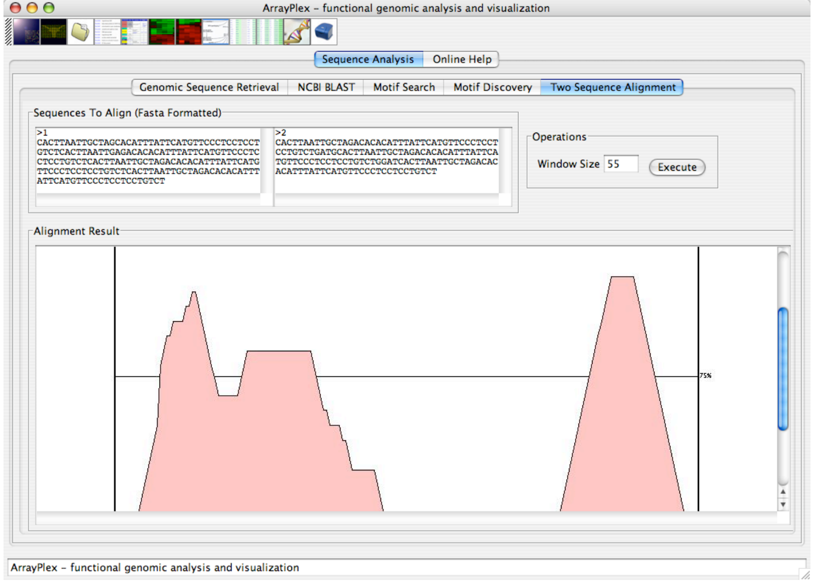
Annotation Format: UID, ORGANISM, CHROMOSOME, STRAND, START, STOP

```
CTCCTGTCTCACTTAATTGCTAGACACACATTTATTCATGTTCCCTCCTCCTGTCTCACTTAATTGCTAGACACACATTTATTCATGTTCCCTCCTCCTGTCTCACTTAATTGCTAGACACACATTTATTCA
TGTTCCCTCCTCCTGTCTCACTTAATTGCTAGACACACATTTATTCATGTTCCCTCCTCCTGTCTCACTTAATTGCTAGACACACATTTATTCATGTTCCCTCCTCCTGTCT
>2
CACTTAATTGCTAGACACACATTTATTCATGTTCCCTCCTCCTGTCTCACTTAATTGCTAGACACACATTTATTCATGTTCCCTCCTCCTGTCTCACTTAATTGCTAGACACACATTTATTCATGTTCCCTC
CTCCTGTCTCACTTAATTGCTAGACACACAT
```

Motif Width 10 # Expected Motifs 10

Motif #	Motif MAP	Sequence ID	Sequence Name	Sequence Positi...	Sequence Strand	Sequence	Max Score (WM)	Score (WM)	Normal (WM)
3	110.172	1	1;	269	0	AATTAAGTG...	22.0	13.1739128...	0.59881422...
3	110.172	1	1;	284	0	AAATGTGTG...	22.0	14.7826087...	0.67193676...
3	110.172	1	1;	330	1	ACTTAATTG...	22.0	11.1739	14.7826087474823 ...
3	110.172	1	1;	331	0	AAATGTGTG...	22.0	14.7826087...	0.67193676...
4	94.5921	0	2;	0	0	GTGTGTCTA...	20.0	20.0	1.0
4	94.5921	0	2;	47	0	GTGTGTCTA...	20.0	20.0	1.0
4	94.5921	0	2;	94	0	GTGTGTCTA...	20.0	20.0	1.0
4	94.5921	0	2;	141	0	GTGTGTCTA...	20.0	20.0	1.0
4	94.5921	1	1;	0	0	GTGTGTCTA...	20.0	20.0	1.0

AI-16 – ArrayPlex Client Sequence Alignment



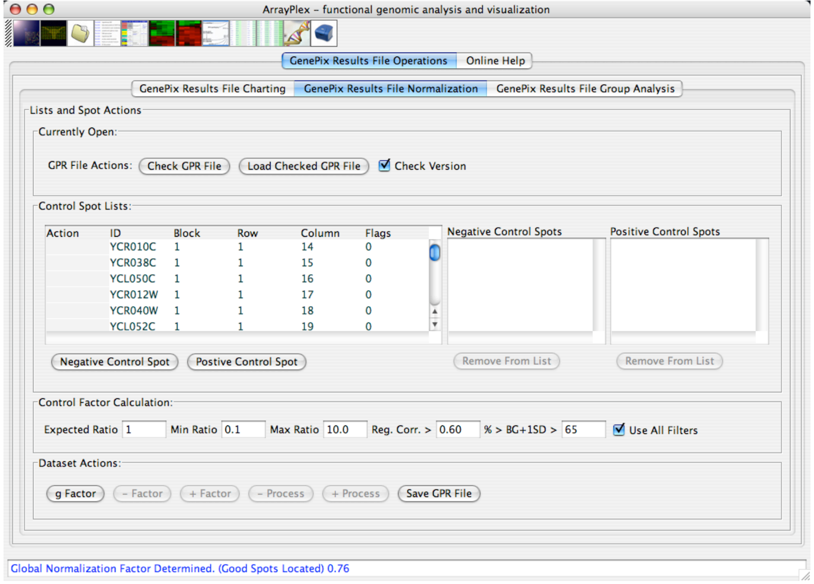
```
>1
CACTTAATTGCTAGCACATTTATTCATGTTCCCTCCTCCT
GTCTCACTTAATTGAGACACACATTTATTCATGTTCCCTC
CTCCTGTCTCACTTAATTGCTAGACACACATTTATTCATG
TTCCCTCCTCCTGTCTCACTTAATTGCTAGACACACATTT
ATTCATGTTCCCTCCTCCTGTCT
```

```
>2
CACTTAATTGCTAGACACACATTTATTCATGTTCCCTCCT
CCTGTCTGATGCACTTAATTGCTAGACACACATTTATTCA
TGTTCCCTCCTCCTGTCTGGATCACTTAATTGCTAGACAC
ACATTTATTCATGTTCCCTCCTCCTGTCT
```

Window Size 55

ArrayPlex – functional genomic analysis and visualization

AI-17 – ArrayPlex Client GenePix Pro Results Normalization



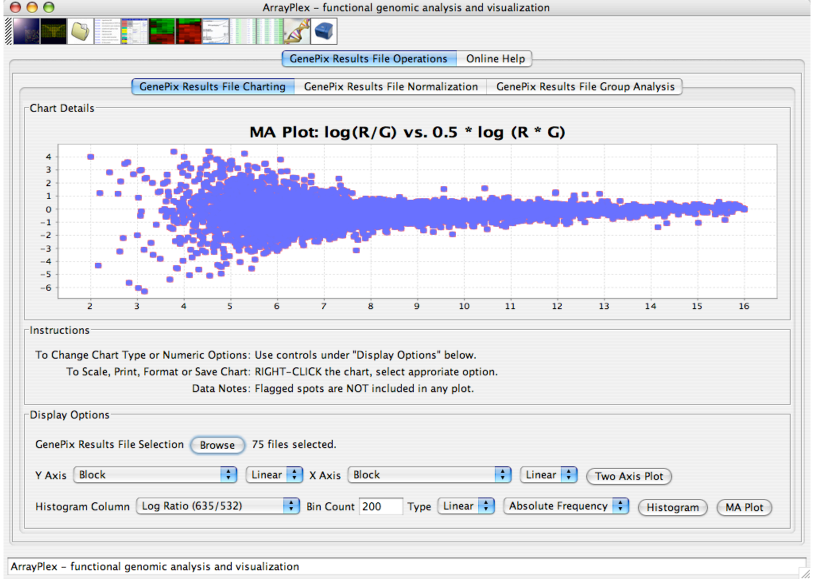
☑ Check Version

Action	ID	Block	Row	Column	Flags
	YCR010C	1	1	14	0
	YCR038C	1	1	15	0
	YCL050C	1	1	16	0
	YCR012W	1	1	17	0
	YCR040W	1	1	18	0
	YCL052C	1	1	19	0

Expected Ratio 1 Min Ratio 0.1 Max Ratio 10.0 Reg. Corr. > 0.60 % > BG+1SD > 65 ☑ Use All Filters

Global Normalization Factor Determined. (Good Spots Located) 0.76

AI-18 – ArrayPlex Client Batch MA Plotting (75 GPR Files)


ArrayPlex – functional genomic analysis and visualization
GenePix Results File Operations
Online Help
GenePix Results File Charting
GenePix Results File Normalization
GenePix Results File Group Analysis
Chart Details
MA Plot: log(R/G) vs. 0.5 * log (R * G)
4
3
2
1
0
-1
-2
-3
-4
-5
-6
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Instructions
Display Options
Browse
Two Axis Plot
Histogram
MA Plot

To Change Chart Type or Numeric Options: Use controls under "Display Options" below.

To Scale, Print, Format or Save Chart: RIGHT-CLICK the chart, select approriate option.

Data Notes: Flagged spots are NOT included in any plot.

GenePix Results File Selection: 75 files selected.

Y Axis: Block, Linear

X Axis: Block, Linear

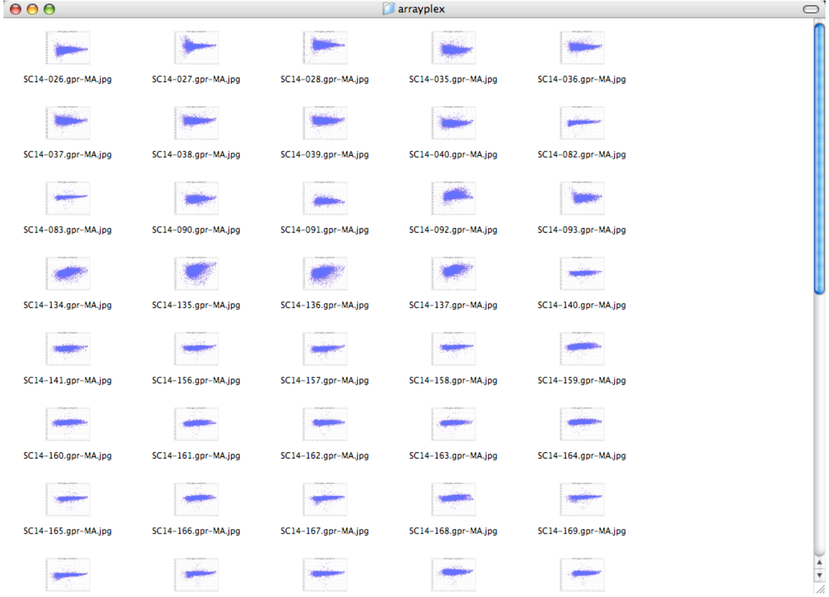
Histogram Column: Log Ratio (635/532)

Bin Count: 200

Type: Linear, Absolute Frequency

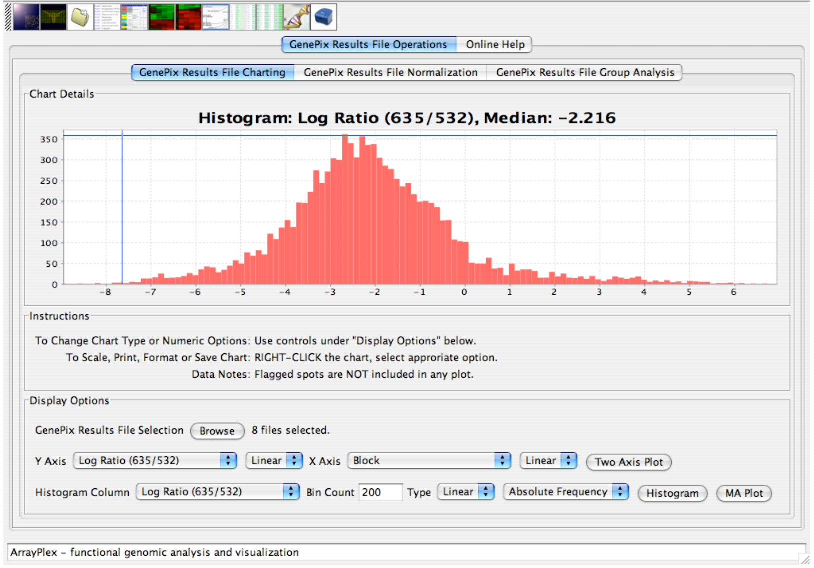
ArrayPlex – functional genomic analysis and visualization

AI-19 – ArrayPlex Client Batch MA Plotting Results



SC14-026.gpr-MA.jpg SC14-027.gpr-MA.jpg SC14-028.gpr-MA.jpg SC14-035.gpr-MA.jpg SC14-036.gpr-MA.jpg
SC14-037.gpr-MA.jpg SC14-038.gpr-MA.jpg SC14-039.gpr-MA.jpg SC14-040.gpr-MA.jpg SC14-082.gpr-MA.jpg
SC14-083.gpr-MA.jpg SC14-090.gpr-MA.jpg SC14-091.gpr-MA.jpg SC14-092.gpr-MA.jpg SC14-093.gpr-MA.jpg
SC14-134.gpr-MA.jpg SC14-135.gpr-MA.jpg SC14-136.gpr-MA.jpg SC14-137.gpr-MA.jpg SC14-140.gpr-MA.jpg
SC14-141.gpr-MA.jpg SC14-156.gpr-MA.jpg SC14-157.gpr-MA.jpg SC14-158.gpr-MA.jpg SC14-159.gpr-MA.jpg
SC14-160.gpr-MA.jpg SC14-161.gpr-MA.jpg SC14-162.gpr-MA.jpg SC14-163.gpr-MA.jpg SC14-164.gpr-MA.jpg
SC14-165.gpr-MA.jpg SC14-166.gpr-MA.jpg SC14-167.gpr-MA.jpg SC14-168.gpr-MA.jpg SC14-169.gpr-MA.jpg

AI-20 – ArrayPlex Client GenePix Pro Results Histogram Plotting


ArrayPlex – functional genomic analysis and visualization
GenePix Results File Operations
Online Help
GenePix Results File Charting
GenePix Results File Normalization
GenePix Results File Group Analysis
Chart Details
350
300
250
200
150
100
50
0
−8
−7
−6
−5
−4
−3
−2
−1
0
1
2
3
4
5
6
Browse
Two Axis Plot
Histogram
MA Plot

Histogram: Log Ratio (635/532), Median: −2.216

Instructions

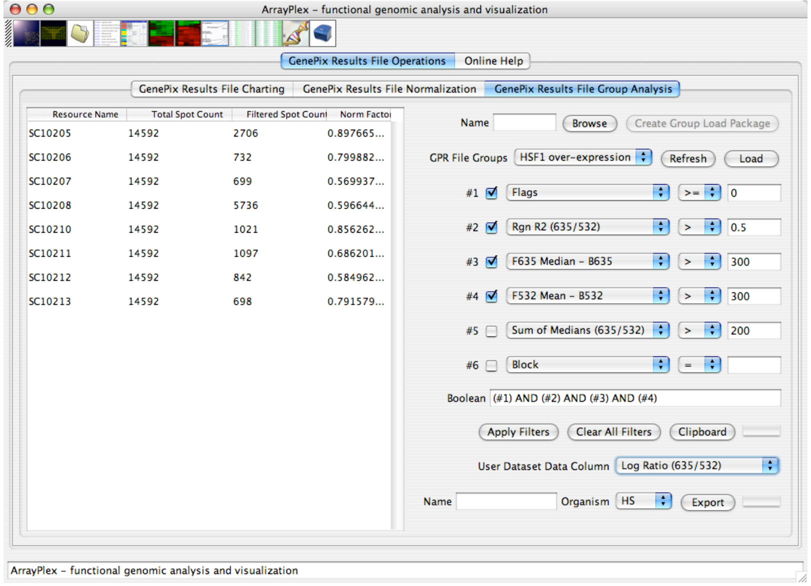
To Change Chart Type or Numeric Options: Use controls under "Display Options" below.
To Scale, Print, Format or Save Chart: RIGHT-CLICK the chart, select approriate option.
Data Notes: Flagged spots are NOT included in any plot.

Display Options

GenePix Results File Selection: 8 files selected.
Y Axis: Log Ratio (635/532), Linear
X Axis: Block, Linear
Histogram Column: Log Ratio (635/532)
Bin Count: 200
Type: Linear
Absolute Frequency

ArrayPlex – functional genomic analysis and visualization

AI-21 – ArrayPlex Client GenePix Pro Results File Group Analysis


ArrayPlex – functional genomic analysis and visualization
GenePix Results File Operations
Online Help
GenePix Results File Charting
GenePix Results File Normalization
GenePix Results File Group Analysis
Browse
Create Group Load Package
Refresh
Load
Apply Filters
Clear All Filters
Clipboard
Export

Resource Name	Total Spot Count	Filtered Spot Count	Norm Factor
SC10205	14592	2706	0.897665...
SC10206	14592	732	0.799882...
SC10207	14592	699	0.569937...
SC10208	14592	5736	0.596644...
SC10210	14592	1021	0.856262...
SC10211	14592	1097	0.686201...
SC10212	14592	842	0.584962...
SC10213	14592	698	0.791579...

Name:
GPR File Groups: HSF1 over-expression

#1 ☑ Flags >= 0
#2 ☑ Rgn R2 (635/532) > 0.5
#3 ☑ F635 Median – B635 > 300
#4 ☑ F532 Mean – B532 > 300
#5 ☐ Sum of Medians (635/532) > 200
#6 ☐ Block =

Boolean: (#1) AND (#2) AND (#3) AND (#4)

User Dataset Data Column: Log Ratio (635/532)

Name:
Organism: HS

ArrayPlex – functional genomic analysis and visualization

Appendix II – GO Slim Term Enrichment

TF Systematic	TF Symbol	GO ID	*P Value*	Term
YBL052C	SAS3	GO:0008372	3.29E-04	cellular component unknown
YBL052C	SAS3	GO:0000004	8.66E-04	biological process unknown
YBL103C	RTG3	GO:0005618	9.54E-05	cell wall
YBR033W		GO:0006519	2.92E-07	amino acid and derivative metabolism
YBR049C	REB1	GO:0005840	2.79E-17	ribosome
YBR049C	REB1	GO:0007582	1.92E-15	physiological process
YBR049C	REB1	GO:0006519	2.06E-08	amino acid and derivative metabolism
YBR049C	REB1	GO:0005198	9.87E-08	structural molecule activity
YBR049C	REB1	GO:0008152	2.18E-06	metabolism
YBR049C	REB1	GO:0006412	2.45E-06	protein biosynthesis
YBR049C	REB1	GO:0043232	9.56E-06	intracellular non-membrane-bound organelle
YBR049C	REB1	GO:0006950	2.71E-05	response to stress
YBR049C	REB1	GO:0043170	1.57E-04	macromolecule metabolism
YBR049C	REB1	GO:0005618	9.50E-04	cell wall
YBR103W	SIF2	GO:0004386	5.67E-04	helicase activity
YBR182C	SMP1	GO:0006810	7.97E-04	transport
YBR240C	THI2	GO:0005840	3.17E-04	ribosome
YBR289W	SNF5	GO:0007582	4.34E-10	physiological process
YBR289W	SNF5	GO:0016829	1.61E-09	lyase activity
YBR289W	SNF5	GO:0006519	2.06E-09	amino acid and derivative metabolism
YBR289W	SNF5	GO:0005886	2.82E-07	plasma membrane
YBR289W	SNF5	GO:0006091	4.21E-05	generation of precursor metabolites and energy
YBR289W	SNF5	GO:0005618	6.13E-05	cell wall
YBR289W	SNF5	GO:0008152	1.69E-04	metabolism
YBR289W	SNF5	GO:0016491	8.30E-04	oxidoreductase activity
YCL055W	KAR4	GO:0000746	9.36E-04	conjugation
YCR081W	SRB8	GO:0005618	9.02E-07	cell wall
YCR081W	SRB8	GO:0006950	6.42E-05	response to stress
YCR084C	TUP1	GO:0005618	3.40E-07	cell wall
YCR084C	TUP1	GO:0007582	4.69E-06	physiological process
YCR084C	TUP1	GO:0016491	1.62E-05	oxidoreductase activity
YCR084C	TUP1	GO:0016829	3.13E-05	lyase activity
YCR106W	RDS1	GO:0006950	4.11E-04	response to stress
YDL020C	RPN4	GO:0005886	1.84E-05	plasma membrane
YDL020C	RPN4	GO:0000746	1.24E-04	conjugation
YDL020C	RPN4	GO:0008233	1.64E-04	peptidase activity
YDL020C	RPN4	GO:0005618	2.54E-04	cell wall
YDL020C	RPN4	GO:0006950	3.51E-04	response to stress
YDL042C	SIR2	GO:0005840	1.72E-09	ribosome
YDL042C	SIR2	GO:0005198	1.31E-06	structural molecule activity
YDL042C	SIR2	GO:0000746	9.36E-06	conjugation
YDL106C	PHO2	GO:0003824	1.92E-05	catalytic activity
YDL106C	PHO2	GO:0005773	1.52E-04	vacuole
YDR073W	SNF11	GO:0005618	4.99E-05	cell wall
YDR146C	SWI5	GO:0005576	3.42E-04	extracellular region
YDR176W	NGG1	GO:0006950	1.21E-04	response to stress
YDR176W	NGG1	GO:0016787	8.40E-04	hydrolase activity
YDR181C	SAS4	GO:0008372	3.57E-06	cellular component unknown
YDR181C	SAS4	GO:0006766	2.48E-04	vitamin metabolism
YDR207C	UME6	GO:0007126	6.97E-08	meiosis
YDR207C	UME6	GO:0006519	4.57E-05	amino acid and derivative metabolism
YDR207C	UME6	GO:0016491	2.65E-04	oxidoreductase activity
YDR259C	YAP6	GO:0005618	4.40E-04	cell wall

YDR310C	SUM1	GO:0030435	2.35E-12	sporulation
YDR310C	SUM1	GO:0005618	1.11E-10	cell wall
YDR310C	SUM1	GO:0005554	3.36E-05	molecular function unknown
YDR310C	SUM1	GO:0008372	5.34E-05	cellular component unknown
YDR310C	SUM1	GO:0000004	3.58E-04	biological process unknown
YDR310C	SUM1	GO:0006766	5.53E-04	vitamin metabolism
YDR363W	ESC2	GO:0006950	6.42E-05	response to stress
YDR392W	SPT3	GO:0006519	3.36E-06	amino acid and derivative metabolism
YDR392W	SPT3	GO:0016491	4.99E-06	oxidoreductase activity
YDR392W	SPT3	GO:0005576	2.67E-05	extracellular region
YDR392W	SPT3	GO:0016829	8.62E-04	lyase activity
YDR392W	SPT3	GO:0003824	9.13E-04	catalytic activity
YDR443C	SSN2	GO:0006950	3.57E-05	response to stress
YDR443C	SSN2	GO:0005618	2.27E-04	cell wall
YDR448W	ADA2	GO:0005618	5.44E-05	cell wall
YDR463W	STP1	GO:0005618	6.72E-06	cell wall
YDR463W	STP1	GO:0006519	2.83E-04	amino acid and derivative metabolism
YDR463W	STP1	GO:0006810	4.82E-04	transport
YDR463W	STP1	GO:0005886	5.77E-04	plasma membrane
YDR477W	SNF1	GO:0005886	3.17E-04	plasma membrane
YDR477W	SNF1	GO:0005215	3.33E-04	transporter activity
YDR520C		GO:0005886	5.71E-04	plasma membrane
YEL009C	GCN4	GO:0006950	6.04E-04	response to stress
YER040W	GLN3	GO:0007582	2.87E-19	physiological process
YER040W	GLN3	GO:0008152	8.56E-15	metabolism
YER040W	GLN3	GO:0016491	2.10E-04	oxidoreductase activity
YER040W	GLN3	GO:0006519	9.21E-04	amino acid and derivative metabolism
YER045C	ACA1	GO:0000746	8.12E-04	conjugation
YER109C	FLO8	GO:0006950	5.99E-05	response to stress
YER109C	FLO8	GO:0005618	3.08E-04	cell wall
YER111C	SWI4	GO:0005618	4.12E-08	cell wall
YER111C	SWI4	GO:0007047	6.25E-04	cell wall organization and biogenesis
YER111C	SWI4	GO:0005576	8.84E-04	extracellular region
YER130C		GO:0006950	4.11E-04	response to stress
YER169W	RPH1	GO:0005618	9.98E-05	cell wall
YER184C		GO:0006091	3.94E-04	generation of precursor metabolites and energy
YFL031W	HAC1	GO:0006810	4.80E-04	transport
YFR034C	PHO4	GO:0005618	3.41E-06	cell wall
YGL025C	PGD1	GO:0006950	1.10E-04	response to stress
YGL071W	RCS1	GO:0005618	4.78E-06	cell wall
YGL071W	RCS1	GO:0016491	7.89E-06	oxidoreductase activity
YGL071W	RCS1	GO:0006950	3.83E-05	response to stress
YGL073W	HSF1	GO:0005840	3.18E-84	ribosome
YGL073W	HSF1	GO:0007582	1.87E-63	physiological process
YGL073W	HSF1	GO:0005198	3.92E-41	structural molecule activity
YGL073W	HSF1	GO:0043232	1.04E-35	intracellular non-membrane-bound organelle
YGL073W	HSF1	GO:0006412	5.32E-31	protein biosynthesis
YGL073W	HSF1	GO:0008152	9.01E-25	metabolism
YGL073W	HSF1	GO:0043170	3.98E-17	macromolecule metabolism
YGL073W	HSF1	GO:0005622	2.71E-15	intracellular
YGL073W	HSF1	GO:0044267	3.24E-15	cellular protein metabolism
YGL073W	HSF1	GO:0019538	7.61E-15	protein metabolism
YGL073W	HSF1	GO:0044238	8.97E-08	primary metabolism
YGL073W	HSF1	GO:0005618	1.88E-06	cell wall
YGL073W	HSF1	GO:0006519	3.40E-05	amino acid and derivative metabolism
YGL073W	HSF1	GO:0005576	5.34E-05	extracellular region
YGL073W	HSF1	GO:0005623	1.07E-04	cell
YGL073W	HSF1	GO:0006091	2.08E-04	generation of precursor metabolites and energy

YGL151W	NUT1	GO:0004386	5.79E-06	helicase activity
YGL151W	NUT1	GO:0005618	8.37E-05	cell wall
YGL166W	CUP2	GO:0006950	9.05E-04	response to stress
YGL237C	HAP2	GO:0005618	1.72E-05	cell wall
YGL237C	HAP2	GO:0000910	4.55E-04	cytokinesis
YGR056W	RSC1	GO:0005618	1.82E-04	cell wall
YGR063C	SPT4	GO:0004386	4.45E-04	helicase activity
YHL020C	OPI1	GO:0006629	8.01E-07	lipid metabolism
YHL020C	OPI1	GO:0003824	5.20E-04	catalytic activity
YHL020C	OPI1	GO:0016740	5.31E-04	transferase activity
YHL025W	SNF6	GO:0007582	5.21E-26	physiological process
YHL025W	SNF6	GO:0005840	8.38E-10	ribosome
YHL025W	SNF6	GO:0008152	7.42E-09	metabolism
YHL025W	SNF6	GO:0005618	1.41E-08	cell wall
YHL025W	SNF6	GO:0016829	8.65E-08	lyase activity
YHL025W	SNF6	GO:0006091	2.19E-07	generation of precursor metabolites and energy
YHL025W	SNF6	GO:0006519	9.93E-07	amino acid and derivative metabolism
YHL025W	SNF6	GO:0045182	5.69E-06	translation regulator activity
YHL025W	SNF6	GO:0005623	9.69E-06	cell
YHL025W	SNF6	GO:0005622	2.66E-05	intracellular
YHL025W	SNF6	GO:0004386	1.33E-04	helicase activity
YHL025W	SNF6	GO:0043232	1.62E-04	intracellular non-membrane-bound organelle
YHL025W	SNF6	GO:0005975	5.10E-04	carbohydrate metabolism
YHL025W	SNF6	GO:0016740	5.70E-04	transferase activity
YHL025W	SNF6	GO:0016874	6.50E-04	ligase activity
YHL025W	SNF6	GO:0042254	6.91E-04	ribosome biogenesis and assembly
YHL027W	RIM101	GO:0005215	1.00E-04	transporter activity
YHL027W	RIM101	GO:0006810	1.28E-04	transport
YHL027W	RIM101	GO:0005886	2.25E-04	plasma membrane
YHL027W	RIM101	GO:0005618	3.95E-04	cell wall
YHR041C	SRB2	GO:0005618	7.61E-05	cell wall
YIL036W	CST6	GO:0005840	8.84E-80	ribosome
YIL036W	CST6	GO:0007582	4.34E-68	physiological process
YIL036W	CST6	GO:0043232	1.75E-31	intracellular non-membrane-bound organelle
YIL036W	CST6	GO:0008152	2.50E-26	metabolism
YIL036W	CST6	GO:0005198	1.01E-25	structural molecule activity
YIL036W	CST6	GO:0006412	9.70E-24	protein biosynthesis
YIL036W	CST6	GO:0043170	1.73E-16	macromolecule metabolism
YIL036W	CST6	GO:0005622	5.64E-16	intracellular
YIL036W	CST6	GO:0044267	1.01E-12	cellular protein metabolism
YIL036W	CST6	GO:0019538	1.06E-12	protein metabolism
YIL036W	CST6	GO:0006091	1.35E-09	generation of precursor metabolites and energy
YIL036W	CST6	GO:0005623	7.11E-07	cell
YIL036W	CST6	GO:0005975	3.80E-06	carbohydrate metabolism
YIL036W	CST6	GO:0045182	6.71E-06	translation regulator activity
YIL036W	CST6	GO:0016740	1.29E-05	transferase activity
YIL036W	CST6	GO:0006519	1.60E-05	amino acid and derivative metabolism
YIL036W	CST6	GO:0044238	3.28E-05	primary metabolism
YIL036W	CST6	GO:0016874	5.90E-05	ligase activity
YIL036W	CST6	GO:0016491	2.62E-04	oxidoreductase activity
YIL036W	CST6	GO:0016829	6.39E-04	lyase activity
YIL036W	CST6	GO:0005618	6.75E-04	cell wall
YIL036W	CST6	GO:0005737	8.62E-04	cytoplasm
YIL084C	SDS3	GO:0000746	5.71E-06	conjugation
YIL084C	SDS3	GO:0005618	1.54E-04	cell wall
YIR033W	MGA2	GO:0006950	5.57E-11	response to stress
YIR033W	MGA2	GO:0016491	1.24E-09	oxidoreductase activity
YIR033W	MGA2	GO:0005618	3.99E-05	cell wall

YIR033W	MGA2	GO:0005576	8.66E-04	extracellular region
YJL056C	ZAP1	GO:0005886	5.88E-05	plasma membrane
YJL110C	GZF3	GO:0005886	6.73E-05	plasma membrane
YJL110C	GZF3	GO:0005215	5.50E-04	transporter activity
YJL127C	SPT10	GO:0005840	9.67E-93	ribosome
YJL127C	SPT10	GO:0005198	8.64E-44	structural molecule activity
YJL127C	SPT10	GO:0043232	9.95E-43	intracellular non-membrane-bound organelle
YJL127C	SPT10	GO:0006412	1.93E-37	protein biosynthesis
YJL127C	SPT10	GO:0007582	4.01E-35	physiological process
YJL127C	SPT10	GO:0044267	5.59E-19	cellular protein metabolism
YJL127C	SPT10	GO:0019538	1.26E-18	protein metabolism
YJL127C	SPT10	GO:0043170	4.04E-16	macromolecule metabolism
YJL127C	SPT10	GO:0005622	2.20E-15	intracellular
YJL127C	SPT10	GO:0008152	3.01E-13	metabolism
YJL127C	SPT10	GO:0045182	2.65E-05	translation regulator activity
YJL127C	SPT10	GO:0044238	2.46E-04	primary metabolism
YJL176C	SWI3	GO:0007582	1.09E-09	physiological process
YJL176C	SWI3	GO:0016491	1.07E-05	oxidoreductase activity
YJL176C	SWI3	GO:0006519	1.67E-05	amino acid and derivative metabolism
YJL176C	SWI3	GO:0005886	2.33E-05	plasma membrane
YJL176C	SWI3	GO:0045182	5.66E-05	translation regulator activity
YJL176C	SWI3	GO:0005618	7.50E-05	cell wall
YJL176C	SWI3	GO:0016829	1.90E-04	lyase activity
YJL176C	SWI3	GO:0016874	5.18E-04	ligase activity
YJR060W	CBF1	GO:0006950	1.52E-07	response to stress
YJR060W	CBF1	GO:0005618	6.20E-06	cell wall
YJR060W	CBF1	GO:0016829	1.44E-04	lyase activity
YJR060W	CBF1	GO:0005886	7.80E-04	plasma membrane
YKL015W	PUT3	GO:0006519	1.90E-04	amino acid and derivative metabolism
YKL032C	IXR1	GO:0005576	2.52E-04	extracellular region
YKL038W	RGT1	GO:0000746	2.52E-05	conjugation
YKL038W	RGT1	GO:0007165	7.07E-05	signal transduction
YKL038W	RGT1	GO:0005576	1.45E-04	extracellular region
YKL043W	PHD1	GO:0000746	3.98E-05	conjugation
YKL043W	PHD1	GO:0007165	1.11E-04	signal transduction
YKL043W	PHD1	GO:0005576	1.95E-04	extracellular region
YKL109W	HAP4	GO:0005576	1.77E-04	extracellular region
YKL112W	ABF1	GO:0016491	6.64E-05	oxidoreductase activity
YKR099W	BAS1	GO:0006519	4.64E-08	amino acid and derivative metabolism
YKR099W	BAS1	GO:0003824	7.46E-07	catalytic activity
YKR099W	BAS1	GO:0016491	3.55E-06	oxidoreductase activity
YKR099W	BAS1	GO:0005622	5.00E-04	intracellular
YKR101W	SIR1	GO:0005576	2.33E-04	extracellular region
YLR013W	GAT3	GO:0006091	7.27E-04	generation of precursor metabolites and energy
YLR039C	RIC1	GO:0005618	2.30E-05	cell wall
YLR039C	RIC1	GO:0006950	6.75E-05	response to stress
YLR039C	RIC1	GO:0016829	4.17E-04	lyase activity
YLR039C	RIC1	GO:0007582	7.22E-04	physiological process
YLR131C	ACE2	GO:0005576	3.03E-06	extracellular region
YLR131C	ACE2	GO:0005618	4.28E-05	cell wall
YLR266C	PDR8	GO:0005886	9.51E-04	plasma membrane
YLR357W	RSC2	GO:0006519	4.42E-04	amino acid and derivative metabolism
YLR357W	RSC2	GO:0007582	4.48E-04	physiological process
YLR357W	RSC2	GO:0008152	5.25E-04	metabolism
YLR403W	SFP1	GO:0005840	1.01E-90	ribosome
YLR403W	SFP1	GO:0007582	8.79E-51	physiological process
YLR403W	SFP1	GO:0043232	2.47E-36	intracellular non-membrane-bound organelle
YLR403W	SFP1	GO:0005198	9.71E-36	structural molecule activity

YLR403W	SFP1	GO:0006412	5.03E-31	protein biosynthesis
YLR403W	SFP1	GO:0008152	2.11E-20	metabolism
YLR403W	SFP1	GO:0043170	8.42E-16	macromolecule metabolism
YLR403W	SFP1	GO:0044267	1.01E-15	cellular protein metabolism
YLR403W	SFP1	GO:0019538	8.90E-15	protein metabolism
YLR403W	SFP1	GO:0005622	4.83E-13	intracellular
YLR403W	SFP1	GO:0045182	3.02E-07	translation regulator activity
YLR403W	SFP1	GO:0005623	5.17E-06	cell
YLR403W	SFP1	GO:0006091	8.12E-06	generation of precursor metabolites and energy
YLR403W	SFP1	GO:0042254	3.68E-05	ribosome biogenesis and assembly
YLR403W	SFP1	GO:0016829	5.55E-05	lyase activity
YLR403W	SFP1	GO:0006519	6.86E-05	amino acid and derivative metabolism
YLR403W	SFP1	GO:0044238	1.09E-04	primary metabolism
YLR418C	CDC73	GO:0000746	9.94E-07	conjugation
YLR418C	CDC73	GO:0005618	3.25E-05	cell wall
YLR418C	CDC73	GO:0005576	8.46E-05	extracellular region
YLR418C	CDC73	GO:0016491	9.28E-05	oxidoreductase activity
YLR418C	CDC73	GO:0004386	5.05E-04	helicase activity
YLR418C	CDC73	GO:0006766	5.06E-04	vitamin metabolism
YLR418C	CDC73	GO:0000004	5.23E-04	biological process unknown
YLR442C	SIR3	GO:0000746	6.31E-07	conjugation
YLR442C	SIR3	GO:0005618	4.58E-04	cell wall
YLR442C	SIR3	GO:0007165	9.45E-04	signal transduction
YLR451W	LEU3	GO:0006519	4.56E-09	amino acid and derivative metabolism
YML007W	YAP1	GO:0006950	1.32E-06	response to stress
YML007W	YAP1	GO:0016491	7.43E-04	oxidoreductase activity
YML051W	GAL80	GO:0005975	1.68E-05	carbohydrate metabolism
YML081W		GO:0005618	8.37E-05	cell wall
YML099C	ARG81	GO:0006519	6.71E-05	amino acid and derivative metabolism
YML102W	CAC2	GO:0006766	2.89E-06	vitamin metabolism
YML102W	CAC2	GO:0016491	7.46E-04	oxidoreductase activity
YMR016C	SOK2	GO:0005886	9.11E-04	plasma membrane
YMR037C	MSN2	GO:0006950	7.59E-06	response to stress
YMR037C	MSN2	GO:0005576	7.89E-04	extracellular region
YMR042W	ARG80	GO:0006519	1.80E-05	amino acid and derivative metabolism
YMR043W	MCM1	GO:0007582	4.37E-18	physiological process
YMR043W	MCM1	GO:0006091	6.08E-09	generation of precursor metabolites and energy
YMR043W	MCM1	GO:0008152	2.64E-08	metabolism
YMR043W	MCM1	GO:0005623	5.21E-08	cell
YMR043W	MCM1	GO:0016491	1.95E-07	oxidoreductase activity
YMR043W	MCM1	GO:0016829	7.64E-04	lyase activity
YMR043W	MCM1	GO:0005618	8.74E-04	cell wall
YMR070W	MOT3	GO:0005618	5.34E-10	cell wall
YMR164C	MSS11	GO:0005886	8.89E-04	plasma membrane
YMR182C	RGM1	GO:0005618	9.89E-05	cell wall
YMR182C	RGM1	GO:0006950	5.48E-04	response to stress
YNL021W	HDA1	GO:0005618	4.59E-06	cell wall
YNL097C	PHO23	GO:0006950	2.53E-08	response to stress
YNL097C	PHO23	GO:0016787	1.33E-05	hydrolase activity
YNL199C	GCR2	GO:0005975	4.35E-15	carbohydrate metabolism
YNL199C	GCR2	GO:0006091	3.58E-13	generation of precursor metabolites and energy
YNL199C	GCR2	GO:0016829	1.06E-06	lyase activity
YNL199C	GCR2	GO:0003824	1.91E-05	catalytic activity
YNL216W	RAP1	GO:0005840	2.59E-87	ribosome
YNL216W	RAP1	GO:0005198	4.37E-41	structural molecule activity
YNL216W	RAP1	GO:0043232	5.52E-32	intracellular non-membrane-bound organelle
YNL216W	RAP1	GO:0006412	2.42E-25	protein biosynthesis
YNL216W	RAP1	GO:0005622	2.33E-14	intracellular

YNL216W	RAP1	GO:0044267	8.61E-12	cellular protein metabolism
YNL216W	RAP1	GO:0006519	5.27E-10	amino acid and derivative metabolism
YNL216W	RAP1	GO:0006091	5.43E-10	generation of precursor metabolites and energy
YNL216W	RAP1	GO:0016491	8.70E-09	oxidoreductase activity
YNL216W	RAP1	GO:0005975	1.79E-06	carbohydrate metabolism
YNL216W	RAP1	GO:0016740	9.29E-05	transferase activity
YNL216W	RAP1	GO:0016829	1.60E-04	lyase activity
YNL216W	RAP1	GO:0005623	2.72E-04	cell
YNL216W	RAP1	GO:0005576	2.73E-04	extracellular region
YNL216W	RAP1	GO:0005618	6.14E-04	cell wall
YNL236W	SIN4	GO:0007582	2.81E-20	physiological process
YNL236W	SIN4	GO:0006091	9.92E-10	generation of precursor metabolites and energy
YNL236W	SIN4	GO:0016829	1.52E-08	lyase activity
YNL236W	SIN4	GO:0008152	1.46E-07	metabolism
YNL236W	SIN4	GO:0006519	7.53E-07	amino acid and derivative metabolism
YNL236W	SIN4	GO:0005618	3.43E-06	cell wall
YNL236W	SIN4	GO:0016491	9.33E-06	oxidoreductase activity
YNL236W	SIN4	GO:0005840	1.62E-04	ribosome
YNL236W	SIN4	GO:0045182	1.98E-04	translation regulator activity
YNL236W	SIN4	GO:0016874	6.71E-04	ligase activity
YNL309W	STB1	GO:0006810	8.05E-05	transport
YNL309W	STB1	GO:0005618	1.32E-04	cell wall
YNR010W	CSE2	GO:0005618	1.46E-04	cell wall
YOL004W	SIN3	GO:0005840	8.56E-09	ribosome
YOL004W	SIN3	GO:0005618	4.48E-05	cell wall
YOL004W	SIN3	GO:0007582	5.35E-05	physiological process
YOL004W	SIN3	GO:0005576	6.51E-05	extracellular region
YOL004W	SIN3	GO:0006519	9.10E-04	amino acid and derivative metabolism
YOL068C	HST1	GO:0006766	1.36E-05	vitamin metabolism
YOL068C	HST1	GO:0030435	6.25E-04	sporulation
YOL068C	HST1	GO:0005618	8.23E-04	cell wall
YOL148C	SPT20	GO:0005840	1.26E-43	ribosome
YOL148C	SPT20	GO:0007582	1.50E-23	physiological process
YOL148C	SPT20	GO:0043232	2.05E-15	intracellular non-membrane-bound organelle
YOL148C	SPT20	GO:0005198	3.53E-15	structural molecule activity
YOL148C	SPT20	GO:0006412	4.13E-11	protein biosynthesis
YOL148C	SPT20	GO:0016829	3.81E-07	lyase activity
YOL148C	SPT20	GO:0005618	5.16E-07	cell wall
YOL148C	SPT20	GO:0008152	5.92E-07	metabolism
YOL148C	SPT20	GO:0006091	3.93E-06	generation of precursor metabolites and energy
YOL148C	SPT20	GO:0006519	8.23E-06	amino acid and derivative metabolism
YOL148C	SPT20	GO:0043170	1.80E-05	macromolecule metabolism
YOL148C	SPT20	GO:0005576	2.84E-05	extracellular region
YOL148C	SPT20	GO:0005886	3.26E-05	plasma membrane
YOL148C	SPT20	GO:0005622	4.90E-05	intracellular
YOL148C	SPT20	GO:0044267	1.72E-04	cellular protein metabolism
YOL148C	SPT20	GO:0045182	1.72E-04	translation regulator activity
YOL148C	SPT20	GO:0019538	3.15E-04	protein metabolism
YOL148C	SPT20	GO:0016491	6.18E-04	oxidoreductase activity
YOR028C	CIN5	GO:0005840	4.53E-07	ribosome
YOR028C	CIN5	GO:0005198	1.08E-04	structural molecule activity
YOR028C	CIN5	GO:0043232	9.75E-04	intracellular non-membrane-bound organelle
YOR032C	HMS1	GO:0005840	7.06E-09	ribosome
YOR032C	HMS1	GO:0005198	1.81E-05	structural molecule activity
YOR032C	HMS1	GO:0043232	2.46E-04	intracellular non-membrane-bound organelle
YOR032C	HMS1	GO:0006412	8.60E-04	protein biosynthesis
YOR140W	SFL1	GO:0005618	1.19E-05	cell wall
YOR140W	SFL1	GO:0006810	2.27E-04	transport

YOR213C	SAS5	GO:0008372	3.81E-04	cellular component unknown
YOR229W	WTM2	GO:0006950	5.28E-04	response to stress
YOR290C	SNF2	GO:0005840	8.64E-24	ribosome
YOR290C	SNF2	GO:0007582	3.32E-23	physiological process
YOR290C	SNF2	GO:0043232	3.15E-11	intracellular non-membrane-bound organelle
YOR290C	SNF2	GO:0042254	3.14E-09	ribosome biogenesis and assembly
YOR290C	SNF2	GO:0006519	6.06E-09	amino acid and derivative metabolism
YOR290C	SNF2	GO:0008152	2.46E-08	metabolism
YOR290C	SNF2	GO:0045182	7.68E-07	translation regulator activity
YOR290C	SNF2	GO:0005198	1.28E-06	structural molecule activity
YOR290C	SNF2	GO:0006412	1.47E-06	protein biosynthesis
YOR290C	SNF2	GO:0016829	3.93E-05	lyase activity
YOR290C	SNF2	GO:0005622	6.72E-05	intracellular
YOR290C	SNF2	GO:0005886	7.63E-05	plasma membrane
YOR290C	SNF2	GO:0006996	7.76E-05	organelle organization and biogenesis
YOR290C	SNF2	GO:0005623	1.44E-04	cell
YOR290C	SNF2	GO:0016491	2.46E-04	oxidoreductase activity
YOR290C	SNF2	GO:0005618	2.49E-04	cell wall
YOR290C	SNF2	GO:0005576	2.79E-04	extracellular region
YOR304W	ISW2	GO:0005840	3.29E-28	ribosome
YOR304W	ISW2	GO:0005198	1.20E-18	structural molecule activity
YOR304W	ISW2	GO:0043232	1.27E-14	intracellular non-membrane-bound organelle
YOR304W	ISW2	GO:0006412	3.74E-11	protein biosynthesis
YOR304W	ISW2	GO:0005622	4.50E-08	intracellular
YOR304W	ISW2	GO:0044267	6.27E-06	cellular protein metabolism
YOR363C	PIP2	GO:0005840	1.06E-09	ribosome
YOR363C	PIP2	GO:0005198	1.45E-07	structural molecule activity
YOR363C	PIP2	GO:0043232	2.20E-05	intracellular non-membrane-bound organelle
YOR363C	PIP2	GO:0006412	2.72E-04	protein biosynthesis
YPL049C	DIG1	GO:0000746	1.80E-08	conjugation
YPL075W	GCR1	GO:0005840	9.37E-74	ribosome
YPL075W	GCR1	GO:0007582	1.70E-71	physiological process
YPL075W	GCR1	GO:0005198	1.21E-35	structural molecule activity
YPL075W	GCR1	GO:0008152	8.75E-26	metabolism
YPL075W	GCR1	GO:0043232	4.46E-25	intracellular non-membrane-bound organelle
YPL075W	GCR1	GO:0005622	7.56E-20	intracellular
YPL075W	GCR1	GO:0006412	8.98E-20	protein biosynthesis
YPL075W	GCR1	GO:0043170	7.25E-13	macromolecule metabolism
YPL075W	GCR1	GO:0044267	9.84E-11	cellular protein metabolism
YPL075W	GCR1	GO:0019538	7.95E-10	protein metabolism
YPL075W	GCR1	GO:0006519	1.20E-09	amino acid and derivative metabolism
YPL075W	GCR1	GO:0005623	2.54E-08	cell
YPL075W	GCR1	GO:0016491	1.03E-07	oxidoreductase activity
YPL075W	GCR1	GO:0006091	8.65E-07	generation of precursor metabolites and energy
YPL075W	GCR1	GO:0016829	1.26E-06	lyase activity
YPL075W	GCR1	GO:0045182	2.57E-06	translation regulator activity
YPL075W	GCR1	GO:0005618	3.42E-06	cell wall
YPL075W	GCR1	GO:0005975	2.10E-05	carbohydrate metabolism
YPL075W	GCR1	GO:0016853	5.91E-05	isomerase activity
YPL075W	GCR1	GO:0044238	6.17E-05	primary metabolism
YPL075W	GCR1	GO:0005783	6.95E-04	endoplasmic reticulum
YPL075W	GCR1	GO:0006629	7.84E-04	lipid metabolism
YPL129W	TAF14	GO:0005576	3.91E-07	extracellular region
YPL129W	TAF14	GO:0000746	1.51E-04	conjugation
YPL139C	UME1	GO:0005618	1.70E-04	cell wall
YPL177C	CUP9	GO:0005886	9.87E-05	plasma membrane
YPL248C	GAL4	GO:0005198	7.42E-06	structural molecule activity
YPL248C	GAL4	GO:0005840	3.43E-05	ribosome

YPL254W	HFI1	GO:0005840	5.67E-18	ribosome
YPL254W	HFI1	GO:0005198	5.09E-09	structural molecule activity
YPL254W	HFI1	GO:0005618	1.64E-08	cell wall
YPL254W	HFI1	GO:0007582	5.20E-08	physiological process
YPL254W	HFI1	GO:0005576	2.05E-06	extracellular region
YPL254W	HFI1	GO:0005886	7.45E-06	plasma membrane
YPL254W	HFI1	GO:0043232	1.33E-04	intracellular non-membrane-bound organelle
YPL254W	HFI1	GO:0016491	1.89E-04	oxidoreductase activity
YPL254W	HFI1	GO:0006412	3.41E-04	protein biosynthesis
YPR018W	RLF2	GO:0006766	4.22E-06	vitamin metabolism
YPR018W	RLF2	GO:0008152	7.50E-04	metabolism
YPR065W	ROX1	GO:0006810	2.13E-05	transport
YPR065W	ROX1	GO:0005618	4.87E-05	cell wall
YPR065W	ROX1	GO:0016491	8.48E-05	oxidoreductase activity
YPR065W	ROX1	GO:0006118	9.06E-05	electron transport
YPR065W	ROX1	GO:0006810	4.88E-04	transport

Appendix III – Novel Sequence Motifs

TF Systematic Name	**TF Gene Name**	**Motif**	***P Value***
YBL021C	HAP3	AACTTTGA	0.000861
YBL052C	SAS3	AGAATTGCAGATT	3.27E-13
YBL052C	SAS3	AGAGGAAGCTG	3.74E-11
YBL052C	SAS3	CCTCGAGGA	6.25E-19
YBL052C	SAS3	GGAATAAAAATC	7.65E-23
YBL052C	SAS3	GGATCAAT	4.8E-21
YBL052C	SAS3	GTTGGGATTCCATTG	0.000137
YBL052C	SAS3	TCGCAGGCCAGAAA	0.000276
YBL052C	SAS3	TCGTCTA	0.000000472
YBL052C	SAS3	TGACGCAAAT	3.56E-15
YBL052C	SAS3	TGGAAGCTGA	6.5E-16
YBL052C	SAS3	TTGCACCAAGGAAGT	0.00000984
YBL103C	RTG3	GAAGCCT	0.000288
YBR049C	REB1	GAAGCTGTCATCG	0.00000027
YBR049C	REB1	GAATTGCAGATTCCC	0.000000014
YBR049C	REB1	GTCTATCAACTAA	0.000000456
YBR049C	REB1	GTGGAAGCTGAA	1.41E-10
YBR049C	REB1	GTTAGAAGATGACGC	9.46E-10
YBR049C	REB1	TGGATTC	0.0000696
YBR083W	TEC1	AATATACTAGAAG	0.000757
YBR083W	TEC1	AATCCTCGAG	1.07E-08
YBR083W	TEC1	ACACCGTATATGA	0.0000651
YBR083W	TEC1	ACATATAAAACG	8.92E-19
YBR083W	TEC1	AGAAATAGTCAT	3.54E-19
YBR083W	TEC1	AGAAGCTGTCATCGA	1.7E-13
YBR083W	TEC1	CACCGTATATGATA	0.0000651
YBR083W	TEC1	CATATAAAATG	4.26E-11
YBR083W	TEC1	CTCCTCGAGGATA	0.000219
YBR083W	TEC1	CTCGAGGAGA	3.1E-21
YBR083W	TEC1	CTGAAGTGCAAGG	0.000000541
YBR083W	TEC1	GAAGCTGA	9.57E-27
YBR083W	TEC1	GAGGAATAATCGTAA	1.02E-12
YBR083W	TEC1	GCTGAAATGCAAGGA	0.000219
YBR083W	TEC1	GCTGTCATCGAAG	1.25E-09
YBR083W	TEC1	GGATCAATGAAT	3.53E-25
YBR083W	TEC1	GTGAGGGTTGAAC	0.00029
YBR083W	TEC1	GTGTAGAATTGCAG	4.34E-13
YBR083W	TEC1	TAGGATCAATGAA	3.53E-25
YBR083W	TEC1	TCCTCAAAATGGAAT	0.000381
YBR083W	TEC1	TCCTCGAGGA	2.89E-23
YBR083W	TEC1	TTAGAAGATGAC	2.55E-19
YBR083W	TEC1	TTAGAGGAAGCTGAA	3.74E-12
YBR083W	TEC1	CCTCGAGG	7.8E-22
YBR083W	TEC1	CACCGTATATG	0.0000798
YBR083W	TEC1	CCTCAAAATGGAAT	0.000381
YBR083W	TEC1	CGTATATGAT	0.00000183
YBR083W	TEC1	CTCAAAATGG	0.0000761

YBR083W	TEC1	CTCAAAATGGAAT	0.000381
YBR083W	TEC1	CTCGAGGA	4.44E-21
YBR083W	TEC1	GGATCAATG	3.12E-24
YBR083W	TEC1	TCCTCGAGG	1.32E-22
YBR083W	TEC1	TTAGAGGAAGC	7.55E-12
YBR103W	SIF2	GGGACTGGCC	0.0000035
YBR150C	TBS1	ATTTCCAGGT	0.00000656
YBR150C	TBS1	GATCCGC	0.0000102
YBR289W	SNF5	ATTAGTGGAAGCT	2.77E-10
YBR289W	SNF5	CGCAAGGATTGA	3.73E-13
YBR289W	SNF5	GAGGAAGCTGAAA	0.000034
YBR289W	SNF5	GTGTAGAATTGCAGA	0.000000465
YBR297W	MAL33	CCGATTG	0.0000657
YBR297W	MAL33	GATGAGCTCA	0.000153
YBR297W	MAL33	TAGATGAGCT	0.0000768
YCR081W	SRB8	CTCTGGA	0.000253
YCR084C	TUP1	AATAAGCTTCTG	0.00000192
YCR084C	TUP1	AATAGGATCA	4.36E-10
YCR084C	TUP1	AGAAATATAGATTCC	0.0000274
YCR084C	TUP1	ATAGAGCTGCTTCAA	2.05E-10
YCR084C	TUP1	CAAGGATTGA	7.12E-11
YCR084C	TUP1	CGCATACGAATACAC	1.31E-11
YCR084C	TUP1	CTGAACGAGGGTC	0.000631
YCR084C	TUP1	GCACTAAAAAA	0.00000106
YCR084C	TUP1	GCTGTCATCG	0.0000137
YCR084C	TUP1	GCTTCTGAACGAGG	0.00000279
YCR084C	TUP1	GTGTTAGAAGATGAC	1.11E-08
YCR084C	TUP1	TGATGAC	0.000977
YCR084C	TUP1	TGTACGAGGGTCC	4.65E-08
YCR084C	TUP1	TTACGTA	0.000305
YCR084C	TUP1	TTCGTTCAAAAACAA	0.000164
YCR084C	TUP1	ATAGAGCTGC	1.05E-11
YCR084C	TUP1	CTGAACGAGGG	0.0000429
YCR084C	TUP1	GCATACGAATAC	1.31E-11
YCR084C	TUP1	GTACGAGGGTCC	4.65E-08
YDL020C	RPN4	GCCAAATTGG	0.000968
YDL042C	SIR2	ACAGAATCTCAAA	0.000000643
YDL042C	SIR2	ACATATAAAATGA	4.01E-08
YDL042C	SIR2	CAGTGACA	0.0000172
YDL042C	SIR2	GAAGCTGTCATC	4.07E-09
YDL042C	SIR2	GAGGTTACTGAG	0.000000643
YDL042C	SIR2	GCCTAAAATAGC	0.000000643
YDL042C	SIR2	GTGGAAGC	4.51E-09
YDL042C	SIR2	GTTGGGATTC	0.00000294
YDL042C	SIR2	TAGGATCAATGA	3.46E-16
YDL042C	SIR2	TGTCACAGGAAA	0.0000201
YDL042C	SIR2	GTCACAGGAA	0.0000038
YDL106C	PHO2	AACAAGGCTC	0.00000191
YDL106C	PHO2	AGGCTCAATGCAT	0.000000365
YDL106C	PHO2	CGTACGA	0.0002
YDL106C	PHO2	GAGTCTT	0.000571
YDL106C	PHO2	GCCCCATAGAGAGC	0.000000365

YDL106C	PHO2	GTGAGAC	0.000111
YDL106C	PHO2	TCACCAG	0.00000116
YDL106C	PHO2	TTTGCGCAACGAA	0.000000365
YDL170W	UGA3	GAGGCTTAT	0.0000719
YDR043C	NRG1	AGGCTTA	0.000684
YDR043C	NRG1	GAGGCTT	0.0000134
YDR043C	NRG1	TGAGGCT	0.000354
YDR073W	SNF11	AACGTATATAAGCT	0.00000993
YDR073W	SNF11	TCACACG	0.00000139
YDR176W	NGG1	AAAGCTGC	0.00000294
YDR176W	NGG1	AAGTCATGAC	6.04E-11
YDR176W	NGG1	AGTCCTC	0.00000201
YDR176W	NGG1	ATAAGCGAGATCT	4.03E-12
YDR176W	NGG1	GACTGAAAG	1.32E-09
YDR176W	NGG1	GCAGTCT	0.00000294
YDR176W	NGG1	GGAGATCT	5.02E-08
YDR176W	NGG1	TAAGCGAGATC	4.03E-12
YDR176W	NGG1	AGCGAGATCT	6.04E-11
YDR181C	SAS4	AGGGTCCAAA	0.000000457
YDR181C	SAS4	GAGGTGC	0.000162
YDR181C	SAS4	GGGTGCAA	0.000963
YDR181C	SAS4	TCGTTCAGAAACAA	0.00000279
YDR207C	UME6	ACCCAGAGGTCAT	7.52E-08
YDR207C	UME6	ACCCGCATTAAAGT	0.00000265
YDR207C	UME6	CACGAGGTTACTGAG	7.51E-08
YDR207C	UME6	GAATCTTCATGTCAG	0.00000265
YDR207C	UME6	GAGCCATTGCATGA	1.68E-09
YDR207C	UME6	TAGCCGC	0.000000242
YDR216W	ADR1	GTTGGCTTGAGA	0.000382
YDR310C	SUM1	ATGTCACAAAA	0.0000146
YDR310C	SUM1	CTGACAC	0.000149
YDR310C	SUM1	GCGTCACAAA	0.000000101
YDR310C	SUM1	GGCGTCAGGA	0.00072
YDR310C	SUM1	GTCACAAA	6.69E-09
YDR310C	SUM1	TGCGTCA	0.000343
YDR310C	SUM1	TTGTGTCACT	0.00072
YDR310C	SUM1	TTTGTGTCAC	0.000196
YDR310C	SUM1	TTTGTGTCAT	0.000736
YDR310C	SUM1	TTTTGTGTCA	0.000017
YDR310C	SUM1	TTGTGAC	0.000000189
YDR310C	SUM1	TGACACA	0.000607
YDR310C	SUM1	ATGTCACAAA	0.0000362
YDR310C	SUM1	CGTCACA	0.00000049
YDR310C	SUM1	GCGTCAG	0.000983
YDR310C	SUM1	GTCACAA	0.00000521
YDR310C	SUM1	TGTGTCA	0.00014
YDR392W	SPT3	CTAGTAT	5.4E-13
YDR392W	SPT3	GGAAGCTG	4.69E-22
YDR392W	SPT3	TGTATACCTAA	2.75E-17
YDR392W	SPT3	TTGTTGGGATTCCA	0.00000217
YDR392W	SPT3	GTTGGGATTCCA	0.00000217
YDR421W	ARO80	TCGTCAT	0.000963

YDR448W	ADA2	AAAGTCTC	0.000351
YDR463W	STP1	AACAGACCTGAGAGC	0.000128
YEL009C	GCN4	CTCAGGT	0.0000321
YER045C	ACA1	AAAAGATGCA	0.00000656
YER051W		TCGTGGA	0.000326
YER111C	SWI4	AAAAAGGGCTCC	4.69E-11
YER111C	SWI4	AGGTACG	0.000616
YER111C	SWI4	ATAGTTAAGATACTG	6.69E-09
YER111C	SWI4	CAAGGAAGTA	0.000000217
YER111C	SWI4	CCTCGAA	0.000287
YER111C	SWI4	CTCGACTAAGCAG	1.54E-10
YER111C	SWI4	GCGCAGATTCTGC	2.37E-11
YER111C	SWI4	GCTAAGCGCAG	0.00000424
YER111C	SWI4	GCTGAGC	0.000726
YER111C	SWI4	GGAAATCTA	0.000123
YER111C	SWI4	GGGACAGACAGTC	2.37E-11
YER111C	SWI4	GGGACAGACAGTCGC	2.37E-11
YER111C	SWI4	TAGGCTAAGC	0.00000631
YER111C	SWI4	TAGTCATACAGACGC	0.000000465
YER111C	SWI4	TGCAGGC	0.000000177
YER111C	SWI4	GACAGACAGTC	2.37E-11
YER111C	SWI4	GACAGTC	0.00000699
YFL021W	GAT1	GGTGCAA	0.000101
YFR034C	PHO4	CTCCCGA	0.0000685
YGL013C	PDR1	AGTTACT	0.000947
YGL025C	PGD1	AAAGCTGCAG	2.2E-09
YGL025C	PGD1	AAATCATGACA	5.27E-11
YGL025C	PGD1	AAGACTC	0.00000251
YGL025C	PGD1	AAGCGAGAT	7.36E-10
YGL025C	PGD1	AATGTACAAGAAC	1.06E-11
YGL025C	PGD1	AGATCTT	0.0000289
YGL025C	PGD1	AGGCTGCTGCCTG	1.06E-11
YGL025C	PGD1	ATTAAAGCTGC	1.06E-11
YGL025C	PGD1	CAAATCATGA	5.27E-11
YGL025C	PGD1	CAGTCTT	0.00000222
YGL025C	PGD1	CATGACACA	3.68E-10
YGL025C	PGD1	CATGACATACA	1.06E-11
YGL025C	PGD1	CATGATGTGC	0.000717
YGL025C	PGD1	CCTCCGAAGG	1.06E-11
YGL025C	PGD1	CGACGAGGAT	5.27E-11
YGL025C	PGD1	CTGCCACGTC	0.000301
YGL025C	PGD1	CTTGGAGAT	3.68E-10
YGL025C	PGD1	GATGCTGTAATCT	1.06E-11
YGL025C	PGD1	GCTATCG	0.00000629
YGL025C	PGD1	GGACGTTCCA	1.06E-11
YGL025C	PGD1	GGAGATCTCG	1.06E-11
YGL025C	PGD1	GGATCTGGCT	0.000301
YGL025C	PGD1	GTCATGA	0.000017
YGL025C	PGD1	GTCGACGAGG	1.06E-11
YGL025C	PGD1	GTCTAAC	0.000000228
YGL025C	PGD1	TATCGCT	0.000197
YGL025C	PGD1	TCAAGCAGCA	7.36E-10

YGL025C	PGD1	TCATGTC	4.39E-08
YGL025C	PGD1	TCATGAT	0.0000303
YGL025C	PGD1	CATGACA	0.000000712
YGL025C	PGD1	TCGACGA	0.00000293
YGL035C	MIG1	GCCCGAT	0.00095
YGL073W	HSF1	CCCATGC	0.0000553
YGL073W	HSF1	CGCACGT	0.0000016
YGL073W	HSF1	GCAAGGATTGA	0.00000215
YGL151W	NUT1	AGTTGAGAGACAGG	0.0000146
YGL151W	NUT1	CTAAGCGCAGG	0.000789
YGL151W	NUT1	GTAGGGTAAC	0.000647
YGL151W	NUT1	TCGCACA	0.000655
YGL166W	CUP2	AATTGAC	0.00000889
YGL181W	GTS1	AACATATAAAATG	3.57E-08
YGL181W	GTS1	AGAATTGCAG	8.01E-09
YGL181W	GTS1	CTAGTATATTATC	4E-12
YGL181W	GTS1	GATGACATAAG	0.0000102
YGL181W	GTS1	GATTCCC	0.00000517
YGL181W	GTS1	TAGTGGAAGCTGAA	3.31E-12
YGL181W	GTS1	TATTATCATATACG	1.21E-15
YGL181W	GTS1	TCTAGTA	9.21E-12
YGL237C	HAP2	AAAAAAATGTATCA	3.11E-08
YGL237C	HAP2	AGGAAGAGCAACGTC	3.11E-08
YGL237C	HAP2	GAGGCCTGAGG	3.11E-08
YGL237C	HAP2	GAGTCTTCAAGCAG	3.11E-08
YGL237C	HAP2	GGATGTTCC	0.00000209
YGL237C	HAP2	TAAGCGAGATCTT	3.11E-08
YGL237C	HAP2	TAAGCGAGATCTTT	3.11E-08
YGL237C	HAP2	TAAGGAAGAGCAAC	3.11E-08
YGL237C	HAP2	TTCGAAAAAAATAGA	3.11E-08
YGL237C	HAP2	AAGCGAGATCT	0.000000154
YGL237C	HAP2	TAAGCGAGATCT	3.11E-08
YGL237C	HAP2	TAAGGAAGAGCA	3.11E-08
YGR040W	KSS1	GGGATTCCAT	0.0000517
YGR040W	KSS1	TGGGATTCCA	0.000047
YGR056W	RSC1	AAGATCTCAGCAGA	0.000256
YGR063C	SPT4	AATCCTCGAGG	0.000000127
YGR063C	SPT4	AGTGGAAGCT	3.73E-12
YGR063C	SPT4	CATCGAAGTTAGAG	0.000505
YGR063C	SPT4	CCTCGAC	0.00000299
YGR063C	SPT4	GAAGATGACGCAAAT	2.35E-11
YGR063C	SPT4	GGCTACGCCG	0.000407
YGR063C	SPT4	TAGATTC	0.000483
YGR063C	SPT4	TGGAATAAAAATC	3.76E-15
YGR089W	NNF2	TTGGGATTCC	0.000546
YHL020C	OPI1	AAAAGACATTTTTG	0.0000385
YHL020C	OPI1	TTCCAGCAAAAA	0.0000385
YHL020C	OPI1	AAAGACATTTTTG	0.0000385
YHL025W	SNF6	AATAGGATCAATGA	0.00000018
YHL025W	SNF6	AGATTCC	0.000277
YHL025W	SNF6	CCTCGAG	0.000152
YHL025W	SNF6	CTGCATAGCGCAG	0.00000167

YHL025W	SNF6	GAGGAATAATCG	0.000525
YHL025W	SNF6	GTGCACCATGGAAAT	0.000011
YHL025W	SNF6	TCCTCGA	4.5E-10
YHL027W	RIM101	CTCGAGG	0.000167
YHR041C	SRB2	AAACCCCGTC	2.36E-09
YHR041C	SRB2	AAATCATGACATA	2.36E-09
YHR041C	SRB2	AAGCGAGATCTTTA	2.36E-09
YHR041C	SRB2	ACAAGGT	0.00091
YHR041C	SRB2	ACTGTAAGATC	1.17E-08
YHR041C	SRB2	AGATCAC	0.000106
YHR041C	SRB2	CTGAGCCGA	3.51E-08
YHR041C	SRB2	CTGCCAAAGG	0.000305
YHR041C	SRB2	CTGTACAAGGCTGC	2.36E-09
YHR041C	SRB2	GACGAGGATGC	2.36E-09
YHR041C	SRB2	GAGATCTCGC	6.77E-10
YHR041C	SRB2	GGCCTGAGGC	2.36E-09
YHR041C	SRB2	TAAAGCTGCAGT	2.36E-09
YHR041C	SRB2	TCAAGCAG	0.00000367
YHR041C	SRB2	TCAGACC	0.000000953
YHR041C	SRB2	AGCGAGATC	3.51E-08
YIL036W	CST6	ACGCAAGGAT	0.000396
YIL036W	CST6	GGTACCG	0.000393
YIL084C	SDS3	AACATATAAAACG	1.62E-14
YIL084C	SDS3	ATCCTTGCGT	0.00000245
YIL084C	SDS3	ATCTACTAACTAGTA	3.16E-15
YIL084C	SDS3	CGCAAGGATTG	1.46E-19
YIL084C	SDS3	CTCGAGGAT	0.000356
YIL084C	SDS3	CTTCTAGTATA	7.17E-21
YIL084C	SDS3	GAAGCTG	1.31E-12
YIL084C	SDS3	GCGAGCGCCT	0.000143
YIL084C	SDS3	TGGCCAG	0.000251
YIL084C	SDS3	TTATCAATCCTTG	0.0000582
YIL084C	SDS3	TCTGGCCAGA	0.000842
YIL084C	SDS3	ATATAAAACG	1.13E-13
YIL084C	SDS3	CGCAAGGATT	1.46E-19
YIR023W	DAL81	CAGCAAAAAAGACT	0.00000569
YJL089W	SIP4	GTGGGTGACC	0.00000261
YJL103C		ATACTAGAAGTTCTC	0.000309
YJL103C		ATATACTAGAAGTT	0.000247
YJL110C	GZF3	AAACAGCGTC	0.00000153
YJL110C	GZF3	AATGCCAA	0.0000949
YJL127C	SPT10	AGTGCCATAAA	1.03E-09
YJL127C	SPT10	GAACGAGGGTCC	0.0000554
YJL127C	SPT10	GAAGATGACG	0.000011
YJL127C	SPT10	GCAAGGATTG	1.02E-10
YJL127C	SPT10	GTGGAAGCTG	0.000000412
YJL127C	SPT10	TACGAAT	0.0000642
YJL127C	SPT10	TCCGTAC	4.08E-18
YJL127C	SPT10	TCGTTCAGAAAC	0.000119
YJL127C	SPT10	TGTGTAGAATTGC	0.0000337
YJL127C	SPT10	TGTTGGGATTCCATT	0.000974
YJL176C	SWI3	ATCCTCGAGGAGA	0.000000315

YJL176C	SWI3	CTAAATTAGTGGA	2.39E-08
YJL176C	SWI3	GAAATAGTCATCTAA	2.39E-08
YJL176C	SWI3	GAGGAAGCTG	0.0000914
YJL176C	SWI3	GGGATTCCATT	0.000459
YJL176C	SWI3	GTATATTATCATATA	1.78E-08
YJL176C	SWI3	TATCCTCGAGGAG	0.000573
YJR140C	HIR3	TTACTTG	0.000981
YKL005C		GCAGTGGC	0.000206
YKL032C	IXR1	AAAATGGAATCTATA	0.000232
YKL032C	IXR1	ACGCAAGGATTG	2.74E-22
YKL032C	IXR1	AGAATTGCAGATTC	1.49E-12
YKL032C	IXR1	AGATGAC	1.06E-08
YKL032C	IXR1	ATATCCTCGAGGA	1.41E-08
YKL032C	IXR1	ATCAATCCTTGCG	0.000184
YKL032C	IXR1	CATATAAAACG	4.75E-15
YKL032C	IXR1	CATATAAAATGATG	1.68E-10
YKL032C	IXR1	CCTCGAGGATATAG	0.000524
YKL032C	IXR1	CTGTCATCGATGT	0.000888
YKL032C	IXR1	GATTCCATTTTGAGG	8.76E-10
YKL032C	IXR1	GTTATATTATCAA	2.7E-09
YKL032C	IXR1	GTTGGGATTCCATT	1.72E-08
YKL032C	IXR1	TAAATCCTCGAGG	5.33E-09
YKL032C	IXR1	TAACACCGTATATG	0.000167
YKL032C	IXR1	TGAGGAATAATCG	9.09E-11
YKL032C	IXR1	TGGATTCCTAA	9.61E-14
YKL032C	IXR1	CATTTTGAGG	2.65E-09
YKL032C	IXR1	CTCGAGGATA	0.000365
YKL112W	ABF1	AAGAAAAATTTTTC	0.0000221
YKL112W	ABF1	AAGATGACGCAAA	0.000000481
YKL112W	ABF1	ATGAGGAATAATC	0.00000279
YKL112W	ABF1	CCACTAATTTAGAT	0.000479
YKL185W	ASH1	CGAAGGTGCC	0.000573
YKR099W	BAS1	AGAAGATGAC	1.07E-14
YKR099W	BAS1	AGGATCAATGAAT	3.16E-20
YKR099W	BAS1	ATTCCATTTTGAG	2.16E-10
YKR099W	BAS1	GACTCCT	0.00000571
YKR099W	BAS1	TACTAGT	1.26E-09
YKR099W	BAS1	TCCTAAATCCTTG	0.000005
YKR099W	BAS1	TGTGTAGAATTGCA	2.05E-11
YKR099W	BAS1	TTAGCGC	0.000948
YKR099W	BAS1	GAAGATGAC	1.26E-12
YKR099W	BAS1	GGATCAATGAA	3.16E-20
YKR099W	BAS1	GTGTAGAATTGCA	2.05E-11
YKR101W	SIR1	AATTAGTGGAAGCT	0.00000842
YKR101W	SIR1	TCCTCGAGGAG	0.000623
YKR101W	SIR1	TTAGTGGAAGCTGAA	0.00000842
YLR014C	PPR1	TGCTGCA	0.000887
YLR014C	PPR1	TGGCCAT	0.000868
YLR176C	RFX1	GTTGCCATGG	0.0000523
YLR176C	RFX1	TTGCCATGGC	0.0000523
YLR176C	RFX1	GTTGCCA	0.0000108
YLR182W	SWI6	ATTCCATTTTGAGGA	0.0000742

YLR182W	SWI6	GCACAGT	0.000234
YLR418C	CDC73	AAAAAGGGCTCCTC	4.21E-13
YLR418C	CDC73	AGTTAAGATACTG	2.59E-11
YLR418C	CDC73	CTGCGCATAC	0.0000171
YLR418C	CDC73	GAAAGTACGTACC	0.000000824
YLR418C	CDC73	GCACCATGGAAAT	2.86E-10
YLR418C	CDC73	GCGCAGATTC	7.08E-10
YLR418C	CDC73	GGACAGACAGTCGC	2.86E-10
YLR418C	CDC73	GTCAAAAAG	0.00000022
YLR418C	CDC73	GTGGACC	0.0000489
YLR418C	CDC73	TGGGTGCA	0.00016
YLR418C	CDC73	CTGCGCA	0.000242
YLR418C	CDC73	GACAGACAGTCGC	2.86E-10
YLR418C	CDC73	GCACCATGG	3.45E-08
YLR442C	SIR3	AAGCTGTCATCGAAG	0.000000203
YLR442C	SIR3	ATTTACGTTACTAGT	2.95E-13
YLR442C	SIR3	GAAACGCAAGGATTG	3.16E-23
YLR442C	SIR3	GTCAGTATGACAAT	0.00000181
YLR442C	SIR3	GTTGTATCTCAAA	4.39E-08
YLR442C	SIR3	TGACATAAGTTATG	9.62E-11
YLR442C	SIR3	GTCAGTATGAC	0.000000144
YLR451W	LEU3	GACTCAG	0.0000721
YLR453C	RIF2	CGTATGC	0.000649
YML081W		AAAAAGCGTAT	0.000275
YML081W		TCCAGCAAAAAAGA	0.00000993
YML102W	CAC2	ACGTATATACATA	0.0000655
YML102W	CAC2	ACTAGTA	0.00000261
YML102W	CAC2	AGGATCAATG	9.65E-10
YML102W	CAC2	AGTGGAAGCTG	1.06E-11
YML102W	CAC2	GTGGAAGCTGAAA	4.75E-11
YML102W	CAC2	TCGTTCA	0.000661
YML102W	CAC2	TTACTAG	0.0000923
YML102W	CAC2	ACGTATA	0.0000771
YMR021C	MAC1	ATATACTAG	0.000014
YMR021C	MAC1	GAAACGCAAGGATT	0.00000131
YMR021C	MAC1	GAACTTCTAGTAT	0.0000752
YMR021C	MAC1	GGAAGCTGAAA	0.00000207
YMR037C	MSN2	AATGTTGGCTCGC	1.58E-11
YMR037C	MSN2	AGATCTCGCT	1.58E-11
YMR037C	MSN2	AGATCTCGCTTA	1.58E-11
YMR037C	MSN2	CGTTCTGAGG	1.58E-11
YMR037C	MSN2	CTTTTCCGAAAGT	1.58E-11
YMR037C	MSN2	GAAAGTCATG	1.58E-11
YMR037C	MSN2	GAATAACGCATAGAG	1.58E-11
YMR037C	MSN2	GACGAGGAT	3.3E-09
YMR037C	MSN2	GGAGATC	0.00000771
YMR037C	MSN2	GGATGCTTTTCCG	1.58E-11
YMR037C	MSN2	GGATGTTCCA	7.91E-11
YMR037C	MSN2	TAAGGAAGAGCAA	1.58E-11
YMR037C	MSN2	TCATGAC	0.000000337
YMR037C	MSN2	GCGAGATC	1.57E-08
YMR037C	MSN2	AGATCTC	0.00000276

YMR037C	MSN2	CGAGATC	0.00000471
YMR037C	MSN2	TAAGGAAGAGC	7.91E-11
YMR070W	MOT3	CAGATAG	0.000541
YMR070W	MOT3	CTCCGAT	0.00014
YMR070W	MOT3	GCAAAAGGGT	0.0000101
YMR070W	MOT3	GGCTCAC	0.000988
YMR075W		TTGCAAA	0.000724
YMR164C	MSS11	AATATCATATAGAAG	0.00000328
YMR164C	MSS11	ACAAGGTTTTGAA	0.0000135
YMR164C	MSS11	CTGTACA	0.0000436
YMR164C	MSS11	TAATCGA	0.0000226
YMR164C	MSS11	ACAAGGTTTTG	0.000045
YMR273C	ZDS1	CCTGGAA	0.000327
YMR273C	ZDS1	TTCCAGG	0.000312
YNL097C	PHO23	AAATCATGAC	2.4E-12
YNL097C	PHO23	AACACATAATG	2.4E-12
YNL097C	PHO23	AAGCGAGATC	2.4E-12
YNL097C	PHO23	AAGTCATGA	1.58E-10
YNL097C	PHO23	ACTTGGA	0.00000155
YNL097C	PHO23	AGATCTCGC	1.01E-10
YNL097C	PHO23	ATAAGCGAG	6.04E-11
YNL097C	PHO23	CATGACATAC	1.68E-11
YNL097C	PHO23	CCGTTCTGAG	2.4E-12
YNL097C	PHO23	CGTCATG	0.000000186
YNL097C	PHO23	CTCCGAAGGG	3.35E-09
YNL097C	PHO23	CTTCAGCACG	7.19E-12
YNL097C	PHO23	GAGGATGCT	1.68E-11
YNL097C	PHO23	GCAGTGTAAACT	4.8E-13
YNL097C	PHO23	GTCATGACAC	7.19E-12
YNL097C	PHO23	TCCGAAGGGT	3.35E-09
YNL097C	PHO23	TCGTCACACAAGG	4.8E-13
YNL097C	PHO23	TGTTAGACTG	3.36E-11
YNL097C	PHO23	TTCGAAAAA	0.000000215
YNL097C	PHO23	CAGTCTA	0.000000741
YNL097C	PHO23	GATCTCG	9.21E-08
YNL097C	PHO23	TCCGAAGG	5.5E-13
YNL199C	GCR2	CTCCACG	0.000974
YNL199C	GCR2	GATATTGAAAGAC	0.000000353
YNL199C	GCR2	GTTACTAGTAT	3.01E-08
YNL199C	GCR2	GTTGGATCTGGAAAG	0.0000513
YNL216W	RAP1	AAATTAGTGGAAGC	0.000261
YNL216W	RAP1	AGAAGATGACGCA	0.000634
YNL216W	RAP1	CACCCGT	0.0000687
YNL216W	RAP1	CCGCTTA	0.000138
YNL216W	RAP1	CGCATACGAATA	0.000255
YNL216W	RAP1	GAAGCTGAAA	0.00000413
YNL236W	SIN4	AATAGGATCAATG	7.15E-08
YNL236W	SIN4	AATTAGTGGAAGCTG	0.00000518
YNL236W	SIN4	AGAGGAAGCTGAA	0.000572
YNL236W	SIN4	CACCATGGAAATTG	0.000143
YNL236W	SIN4	CACGTAA	0.0000836
YNL236W	SIN4	GATGATGACATAAG	0.000139

YNL236W	SIN4	GATTCCT	0.0000363
YNL236W	SIN4	GGATCAATGA	0.00000018
YNL236W	SIN4	GTTAGAGGAAGCTG	0.000376
YNL309W	STB1	GGTGCAAAAA	0.000813
YNL309W	STB1	GTGCAAAAAAA	0.000652
YOL004W	SIN3	ACGATTATCGAGT	2.18E-09
YOL004W	SIN3	ATTACGATTATCGAG	2.18E-09
YOL004W	SIN3	CCAGAGGTCATGC	0.000000249
YOL068C	HST1	AGTCACTGTCAAGAG	0.000000292
YOL068C	HST1	CTGTCAGTCA	0.00000289
YOL068C	HST1	GTCAGTCACT	0.00001
YOL068C	HST1	TGTCATT	0.0000245
YOL108C	INO4	CAAGTTG	0.000139
YOL116W	MSN1	AGATCAG	0.000284
YOL148C	SPT20	AGTGGAAGC	0.00000225
YOL148C	SPT20	GATGATGACAT	0.0000447
YOL148C	SPT20	TCGAGGA	0.0000242
YOR028C	CIN5	CCAAGTT	0.000895
YOR191W	RIS1	GAGGAAGCT	0.00000596
YOR191W	RIS1	GGAAGCT	0.000107
YOR213C	SAS5	AACGAAT	0.000294
YOR213C	SAS5	AATTCAAGAG	0.0000424
YOR213C	SAS5	ACATACG	0.00000104
YOR213C	SAS5	ACGAATCGTT	8.51E-10
YOR213C	SAS5	ACGACTC	0.00000954
YOR213C	SAS5	AGCTGCT	9.13E-08
YOR213C	SAS5	CGCATACGAA	1.66E-09
YOR213C	SAS5	CTGTACGAGGGTCC	6.89E-08
YOR213C	SAS5	GAATCGT	0.0000187
YOR213C	SAS5	GAGCGTCTGT	0.000000138
YOR213C	SAS5	GAGCTGC	6.64E-09
YOR213C	SAS5	GAGTGCC	0.0001
YOR213C	SAS5	GCGCATACGA	5.86E-10
YOR213C	SAS5	GTATACG	0.0000201
YOR213C	SAS5	GTGCCATAAA	0.000000181
YOR213C	SAS5	TAGAGCTG	1.12E-09
YOR213C	SAS5	TATACGA	0.000073
YOR213C	SAS5	TGCGCATACG	9.09E-11
YOR213C	SAS5	TTCGTTCAAAAAC	0.0000296
YOR213C	SAS5	GATTCGT	0.0000428
YOR213C	SAS5	GCATACG	0.0000132
YOR229W	WTM2	TCCCGAG	0.000982
YOR290C	SNF2	AGTGGAAGCTGAA	0.000000104
YOR290C	SNF2	CTGTCATCGAAGTTA	0.000177
YOR290C	SNF2	GGATTCC	4.31E-09
YOR290C	SNF2	TATCCTCGAGGAGA	0.000266
YOR290C	SNF2	TTCGATG	0.000952
YOR304W	ISW2	AAAAAAGTTCCTG	0.000136
YOR304W	ISW2	ATCCGTC	0.000661
YOR304W	ISW2	CGTACAT	0.000023
YOR304W	ISW2	CTGGCCA	0.000594
YOR344C	TYE7	AACTTCTAGTA	2.57E-17

YOR344C	TYE7	ATAGGATCAATGA	2.03E-14
YOR344C	TYE7	ATGGATTCCTAA	1.77E-09
YOR344C	TYE7	CCTCGAGGAG	3.46E-12
YOR344C	TYE7	GAAATAGTCATC	1.28E-10
YOR344C	TYE7	GACACGT	0.000519
YOR344C	TYE7	GAGGAAGCTGA	3.23E-08
YOR344C	TYE7	GGAAGCTGAA	3.02E-18
YOR344C	TYE7	GTTGGAATAAAAATC	4.48E-14
YOR344C	TYE7	TCCACGC	0.000081
YOR344C	TYE7	TTAGTGGAAGCTG	6.13E-11
YOR358W	HAP5	TTTCGAG	0.000164
YPL049C	DIG1	AAAACGTATATAAGC	2.41E-08
YPL049C	DIG1	ACGTGGG	0.000672
YPL049C	DIG1	TGCCAAA	0.0000193
YPL049C	DIG1	AACGTATATAAGC	6.16E-11
YPL075W	GCR1	CCGTACA	2.46E-14
YPL075W	GCR1	GAAAAATTTTC	0.000542
YPL129W	TAF14	ACGTATATAAA	0.0000161
YPL177C	CUP9	GAAGCTT	0.000462
YPL248C	GAL4	ATCCGTG	0.0000284
YPL254W	HFI1	ACCACGT	0.0000143
YPL254W	HFI1	ACGCAAGGATT	2.61E-12
YPL254W	HFI1	ATATACG	3.54E-09
YPL254W	HFI1	ATTTGGC	0.000173
YPL254W	HFI1	GAAGATGACGC	1.41E-08
YPL254W	HFI1	GTGTAGAATTGC	0.00000162
YPL254W	HFI1	TACGTAA	0.000172
YPL254W	HFI1	TAGTGGAAGCTG	1.2E-09
YPL254W	HFI1	TCCACGT	0.0000438
YPL254W	HFI1	TGTTGGGATTCC	0.000378
YPR018W	RLF2	AAATCCTCGAGGAG	1.42E-09
YPR018W	RLF2	ATAGGATCAATGAA	5E-10
YPR018W	RLF2	GATCAATGAAT	6.24E-10
YPR018W	RLF2	AAATCCTCGAGGA	1.79E-09
YPR054W	SMK1	GCCGCAT	0.000523
YPR054W	SMK1	TGCGCAT	0.000736
YPR065W	ROX1	GGATGCA	0.00074
YPR065W	ROX1	GGGTGCA	0.0000563

References

1. Hodgkin, J., *What does a worm want with 20,000 genes?* Genome Biol, 2001. **2**(11): p. COMMENT2008.
2. Stein, L., et al., *WormBase: network access to the genome and biology of Caenorhabditis elegans*. Nucleic Acids Res, 2001. **29**(1): p. 82-6.
3. Adams, M.D., et al., *The genome sequence of Drosophila melanogaster*. Science, 2000. **287**(5461): p. 2185-95.
4. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
5. Graveley, B.R., *Alternative splicing: increasing diversity in the proteomic world*. Trends Genet, 2001. **17**(2): p. 100-7.
6. Johnson, J.M., et al., *Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays*. Science, 2003. **302**(5653): p. 2141-4.
7. Lee, T.I. and R.A. Young, *Transcription of eukaryotic protein-coding genes*. Annu Rev Genet, 2000. **34**: p. 77-137.
8. Levine, M. and R. Tjian, *Transcription regulation and animal diversity*. Nature, 2003. **424**(6945): p. 147-51.
9. Kornberg, R.D., *Eukaryotic transcriptional control*. Trends Cell Biol, 1999. **9**(12): p. M46-9.
10. Wray, G.A., et al., *The evolution of transcriptional regulation in eukaryotes*. Mol Biol Evol, 2003. **20**(9): p. 1377-419.
11. Marrack, P., et al., *Genomic-scale analysis of gene expression in resting and activated T cells*. Curr Opin Immunol, 2000. **12**(2): p. 206-9.
12. Glynne, R.J., G. Ghandour, and C.C. Goodnow, *Genomic-scale gene expression analysis of lymphocyte growth, tolerance and malignancy*. Curr Opin Immunol, 2000. **12**(2): p. 210-4.
13. Alizadeh, A.A. and L.M. Staudt, *Genomic-scale gene expression profiling of normal and malignant immune cells*. Curr Opin Immunol, 2000. **12**(2): p. 219-25.
14. Sudarsanam, P., et al., *Whole-genome expression analysis of snf/swi mutants of Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. U.S.A., 2000. **97**(7): p. 3364-9.
15. Segal, E., et al., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data*. Nat Genet, 2003. **34**(2): p. 166-76.
16. Yang, S.H., A.D. Sharrocks, and A.J. Whitmarsh, *Transcriptional regulation by the MAP kinase signaling cascades*. Gene, 2003. **320**: p. 3-21.

17. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14*. Cell, 1993. **75**(5): p. 843-54.
18. Nelson, P., et al., *The microRNA world: small is mighty*. Trends Biochem Sci, 2003. **28**(10): p. 534-40.
19. Reinhart, B.J., et al., *The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans*. Nature, 2000. **403**(6772): p. 901-6.
20. He, L. and G.J. Hannon, *MicroRNAs: small RNAs with a big role in gene regulation*. Nat Rev Genet, 2004. **5**(7): p. 522-31.
21. Lund, E., et al., *Nuclear export of microRNA precursors*. Science, 2004. **303**(5654): p. 95-8.
22. Ketting, R.F., et al., *Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans*. Genes Dev, 2001. **15**(20): p. 2654-9.
23. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function*. Cell, 2004. **116**(2): p. 281-97.
24. Lau, N.C., et al., *An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans*. Science, 2001. **294**(5543): p. 858-62.
25. Lim, L.P., et al., *The microRNAs of Caenorhabditis elegans*. Genes Dev, 2003. **17**(8): p. 991-1008.
26. Johnson, S.M., S.Y. Lin, and F.J. Slack, *The time of appearance of the C. elegans let-7 microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter*. Dev Biol, 2003. **259**(2): p. 364-79.
27. Sempere, L.F., et al., *Temporal regulation of microRNA expression in Drosophila melanogaster mediated by hormonal signals and broad-Complex gene activity*. Dev Biol, 2003. **259**(1): p. 9-18.
28. Zhou, X., et al., *Characterization and Identification of MicroRNA Core Promoters in Four Model Species*. PLoS Comput Biol, 2007. **3**(3): p. e37.
29. Shalgi, R., et al., *Global and Local Architecture of the Mammalian microRNA-Transcription Factor Regulatory Network*. PLoS Comput Biol, 2007. **3**(7): p. e131.
30. Ozen, M., et al., *Widespread deregulation of microRNA expression in human prostate cancer*. Oncogene, 2007.
31. Novotny, G.W., et al., *Translational repression of E2F1 mRNA in carcinoma in situ and normal testis correlates with expression of the miR-17-92 cluster*. Cell Death Differ, 2007. **14**(4): p. 879-82.
32. Fontana, L., et al., *MicroRNAs 17-5p-20a-106a control monocytopoiesis through AML1 targeting and M-CSF receptor upregulation*. Nat Cell Biol, 2007.

33. Corney, D.C., et al., *MicroRNA-34b and MicroRNA-34c Are Targets of p53 and Cooperate in Control of Cell Proliferation and Adhesion-Independent Growth*. Cancer Res, 2007.

34. Coller, H.A., J.J. Forman, and A. Legesse-Miller, *"Myc'ed messages": myc induces transcription of E2F1 while inhibiting its translation via a microRNA polycistron*. PLoS Genet, 2007. **3**(8): p. e146.

35. Chang, T.C., et al., *Transactivation of miR-34a by p53 Broadly Influences Gene Expression and Promotes Apoptosis*. Mol. Cell, 2007. **26**(5): p. 745-52.

36. Woods, K., J.M. Thomson, and S.M. Hammond, *Direct regulation of an oncogenic micro-RNA cluster by E2F transcription factors*. J. Biol. Chem., 2006. **282**(4): p. 2130-4.

37. Sylvestre, Y., et al., *An E2F/miR-20a autoregulatory feedback loop*. J. Biol. Chem., 2006. **282**(4): p. 2135-43.

38. Hossain, A., M.T. Kuo, and G.F. Saunders, *Mir-17-5p regulates breast cancer cell proliferation by inhibiting translation of AIB1 mRNA*. Mol. Cell. Biol., 2006. **26**(21): p. 8191-201.

39. O'Donnell, K.A., et al., *c-Myc-regulated microRNAs modulate E2F1 expression*. Nature, 2005. **435**(7043): p. 839-43.

40. Killion, P.J., G. Sherlock, and V.R. Iyer, *The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD)*. BMC Bioinformatics, 2003. **4**: p. 32.

41. Killion, P.J. and V.R. Iyer, *Microarray Data Visualization and Analysis with the Longhorn Array Database (LAD)*. Current Protocols in Bioinformatics, 2004.

42. Saal, L.H., et al., *BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data*. Genome Biol., 2002. **3**(8): p. SOFTWARE0003.

43. Saeed, A.I., et al., *TM4: a free, open-source system for microarray data management and analysis*. BioTechniques, 2003. **34**(2): p. 374-8.

44. Sherlock, G., et al., *The Stanford Microarray Database*. Nucleic Acids Res, 2001. **29**(1): p. 152-5.

45. Brazma, A., et al., *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data*. Nat Genet, 2001. **29**(4): p. 365-71.

46. Hu, Z., P.J. Killion, and V.R. Iyer, *Genetic reconstruction of a functional transcriptional regulatory network*. Nat. Genet., 2007.

47. Harbison, C.T., et al., *Transcriptional regulatory code of a eukaryotic genome*. Nature, 2004. **431**(7004): p. 99-104.

48. Lee, T.I., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae*. Science, 2002. **298**(5594): p. 799-804.

49. Winzeler, E.A., et al., *Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis*. Science, 1999. **285**(5429): p. 901-6.
50. Lieb, J.D., et al., *Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association*. Nat. Genet., 2001. **28**(4): p. 327-34.
51. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks*. Methods, 2003. **13**(11): p. 2498-504.
52. Lydall, D., G. Ammerer, and K. Nasmyth, *A new role for MCM1 in yeast: cell cycle regulation of SW15 transcription*. Genes Dev, 1991. **5**(12B): p. 2405-19.
53. Gimeno, C.J. and G.R. Fink, *Induction of pseudohyphal growth by overexpression of PHD1, a Saccharomyces cerevisiae gene related to transcriptional regulators of fungal development*. Mol. Cell. Biol., 1994. **14**(3): p. 2100-12.
54. Lorenz, M.C. and J. Heitman, *Regulators of pseudohyphal differentiation in Saccharomyces cerevisiae identified through multicopy suppressor analysis in ammonium permease mutant strains*. Genetics, 1998. **150**(4): p. 1443-57.
55. Van Driessche, N., et al., *Epistasis analysis with global transcriptional phenotypes*. Nat. Genet., 2005. **37**(5): p. 471-7.
56. Kundaje, A., et al., *Learning regulatory programs that accurately predict differential expression with MEDUSA*. Ann N Y Acad Sci, 2007.
57. Bailey, T.L. and W.S. Noble, *Searching for statistically significant regulatory modules*. Bioinformatics, 2003. **19 Suppl 2**: p. ii16-25.
58. Yeang, C.H., et al., *Validation and refinement of gene-regulatory pathways on a network of physical interactions*. Genome Biol., 2005. **6**(7): p. R62.
59. Hartemink, A.J., et al., *Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks*. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 2001: p. 422-33.
60. Doniger, S.W. and J.C. Fay, *Frequent gain and loss of functional transcription factor binding sites*. PLoS Comput Biol, 2007. **3**(5): p. e99.
61. Gu, J. and V.R. Iyer, *PI3K signaling and miRNA expression during the response of quiescent human fibroblasts to distinct proliferative stimuli*. Genome Biol., 2006. **7**(5): p. R42.
62. Yoon, S. and G. De Micheli, *Prediction and Analysis of Human microRNA Regulatory Modules*. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference, 2007. **5**: p. 4799-802.
63. Xu, X., *Same computational analysis, different miRNA target predictions*. Nat. Methods, 2007. **4**(3): p. 191; author reply 191.
64. Tian, F., et al., *miRAS: a data processing system for miRNA expression profiling study*. BMC Bioinformatics, 2007. **8**(1): p. 285.
65. Nielsen, C.B., et al., *Determinants of targeting by endogenous and exogenous microRNAs and siRNAs*. RNA, 2007. **13**(11): p. 1894-910.

66. Mazière, P. and A.J. Enright, *Prediction of microRNA targets*. Drug Discov Today, 2007. **12**(11-12): p. 452-8.

67. Lindow, M. and J. Gorodkin, *Principles and Limitations of Computational MicroRNA Gene and Target Finding*. DNA Cell Biol, 2007. **26**(5): p. 339-51.

68. Lee, J., et al., *Regulatory Circuit of Human MicroRNA Biogenesis*. PLoS Comput Biol, 2007. **3**(4): p. e67.

69. John, B., C. Sander, and D.S. Marks, *Prediction of human microRNA targets*. Methods Mol Biol, 2006. **342**: p. 101-13.

70. Yoon, S. and G. De Micheli, *Prediction of regulatory modules comprising microRNAs and target genes*. Bioinformatics, 2005. **21 Suppl 2**: p. ii93-ii100.

71. Dang, C.V., et al., *The c-Myc target gene network*. Semin Cancer Biol, 2006. **16**(4): p. 253-64.

72. Patel, J.H., et al., *Analysis of genomic targets reveals complex functions of MYC*. Nat Rev Cancer, 2004. **4**(7): p. 562-8.

73. Wang, Y., et al., *MicroRNA: past and present*. Front Biosci, 2007. **12**: p. 2316-29.

74. Zhang, B., Q. Wang, and X. Pan, *MicroRNAs and their regulatory roles in animals and plants*. J Cell Physiol, 2007. **210**(2): p. 279-89.

75. Smyth, G.K. and T. Speed, *Normalization of cDNA microarray data*. Methods, 2003. **31**(4): p. 265-73.

76. He, L., et al., *A microRNA component of the p53 tumour suppressor network*. Nature, 2007.

77. Tarasov, V., et al., *Differential Regulation of microRNAs by p53 Revealed by Massively Parallel Sequencing: miR-34a is a p53 Target That Induces Apoptosis and G(1)-arrest*. Cell Cycle, 2007. **6**(13).

78. Raver-Shapira, N., et al., *Transcriptional Activation of miR-34a Contributes to p53-Mediated Apoptosis*. Mol. Cell, 2007. **26**(5): p. 731-43.

79. Mnjoyan, Z.H., et al., *Paradoxical upregulation of tumor suppressor protein p53 in serum-stimulated vascular smooth muscle cells: a novel negative-feedback regulatory mechanism*. Circulation, 2003. **108**(4): p. 464-71.

80. Kent, W.J., et al., *The human genome browser at UCSC*. Methods, 2002. **12**(6): p. 996-1006.

81. Loots, G.G., et al., *rVista for comparative sequence-based discovery of functional transcription factor binding sites*. Methods, 2002. **12**(5): p. 832-9.

82. Griffiths-Jones, S., *miRBase: the microRNA sequence database*. Methods Mol Biol, 2006. **342**: p. 129-38.

83. Weber, M.J., *New human and mouse microRNA genes found by homology search*. FEBS J, 2005. **272**(1): p. 59-73.

84. Meng, F., et al., *Involvement of human micro-RNA in growth and response to chemotherapy in human cholangiocarcinoma cell lines*. Gastroenterology, 2006. **130**(7): p. 2113-29.

85. Walhout, A.J., P.C. van der Vliet, and H.T. Timmers, *Sequences flanking the E-box contribute to cooperative binding by c-Myc/Max heterodimers to adjacent binding sites*. Biochim Biophys Acta, 1998. **1397**(2): p. 189-201.
86. Desbarats, L., S. Gaubatz, and M. Eilers, *Discrimination between different E-box-binding proteins at an endogenous target gene of c-myc*. Genes Dev, 1996. **10**(4): p. 447-60.
87. Bruno, M.E., et al., *Upstream stimulatory factor but not c-Myc enhances transcription of the human polymeric immunoglobulin receptor gene*. Mol Immunol, 2004. **40**(10): p. 695-708.
88. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes*. Methods, 2005. **15**(8): p. 1034-50.
89. Fernandez, P.C., et al., *Genomic targets of the human c-Myc protein*. Genes Dev, 2003. **17**(9): p. 1115-29.
90. Zeller, K.I., et al., *Global mapping of c-Myc binding sites and target gene networks in human B cells*. Proc. Natl. Acad. Sci. U.S.A., 2006. **103**(47): p. 17834-9.
91. Weber, B., et al., *Methylation of Human MicroRNA Genes in Normal and Neoplastic Cells*. Cell Cycle, 2007. **6**(9).
92. Mao, D.Y., et al., *Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression*. Curr Biol, 2003. **13**(10): p. 882-6.
93. Cawley, S., et al., *Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs*. Cell, 2004. **116**(4): p. 499-509.
94. Zeller, K.I., et al., *An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets*. Genome Biol., 2003. **4**(10): p. R69.
95. Slack, A.D., et al., *MYCN-directed centrosome amplification requires MDM2-mediated suppression of p53 activity in neuroblastoma cells*. Cancer Res, 2007. **67**(6): p. 2448-55.
96. Bindra, R.S. and P.M. Glazer, *Co-repression of mismatch repair gene expression by hypoxia in cancer cells: role of the Myc/Max network*. Cancer Lett, 2007. **252**(1): p. 93-103.
97. Bil'din, V.N., T.B. Seregina, and T.V. Pospelova, *[The regulation of DNA repair processes in mammalian cells. II. The repair of DNA radiation damage in NIH 3T3 murine cells transformed by the v-myc oncogene]*. Tsitologiia, 1991. **33**(5): p. 39-47.
98. Jin, Z., et al., *Bcl2 suppresses DNA repair by enhancing c-Myc transcriptional activity*. J. Biol. Chem., 2006. **281**(20): p. 14446-56.

99. Beecham, E.J., et al., *DNA repair in the c-myc proto-oncogene locus: possible involvement in susceptibility or resistance to plasmacytoma induction in BALB/c mice*. Mol. Cell. Biol., 1991. **11**(6): p. 3095-104.
100. Coulouarn, C., et al., *Oncogene-specific gene expression signatures at preneoplastic stage in mice define distinct mechanisms of hepatocarcinogenesis*. Hepatology, 2006. **44**(4): p. 1003-11.
101. Wade, M. and G.M. Wahl, *c-Myc, genome instability, and tumorigenesis: the devil is in the details*. Curr Top Microbiol Immunol, 2006. **302**: p. 169-203.
102. Sethupathy, P., M. Megraw, and A.G. Hatzigeorgiou, *A guide through present computational approaches for the identification of mammalian microRNA targets*. Nat. Methods, 2006. **3**(11): p. 881-6.
103. Hayashi, H., et al., *High-resolution mapping of DNA methylation in human genome using oligonucleotide tiling array*. Hum Genet, 2007. **120**(5): p. 701-11.
104. Consortium, E.P., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project*. Nature, 2007. **447**(7146): p. 799-816.
105. Xie, X., et al., *Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites*. Proc. Natl. Acad. Sci. U.S.A., 2007. **104**(17): p. 7145-50.
106. 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.
107. Shivaswamy, S. and V.R. Iyer, *Genome-wide analysis of chromatin status using tiling microarrays*. Methods, 2007. **41**(3): p. 304-11.
108. Lin, J.M., et al., *Transcription factor binding and modified histones in human bidirectional promoters*. Methods, 2007. **17**(6): p. 818-27.
109. Smalheiser, N.R. and V.I. Torvik, *Complications in mammalian microRNA target prediction*. Methods Mol Biol, 2006. **342**: p. 115-27.
110. Gaidatzis, D., et al., *Inference of miRNA targets using evolutionary conservation and pathway analysis*. BMC Bioinformatics, 2007. **8**: p. 69.
111. Turner, B.M., *Histone acetylation and an epigenetic code*. Bioessays, 2000. **22**(9): p. 836-45.
112. Ruby, J.G., C.H. Jan, and D.P. Bartel, *Intronic microRNA precursors that bypass Drosha processing*. Nature, 2007. **448**(7149): p. 83-6.
113. Kim, Y.K. and V.N. Kim, *Processing of intronic microRNAs*. EMBO J., 2007. **26**(3): p. 775-83.
114. Lee, Y., et al., *Drosha in primary microRNA processing*. Cold Spring Harb Symp Quant Biol, 2006. **71**: p. 51-7.

Vita

Patrick Killion was born on the morning of April 19, 1974 in Oklahoma City, Oklahoma to Jerald and Cathy Killion. He attended Texas A&M University and earned a B.S. in Computer Science in 1997. He was professionally employed as a software engineer for Tivoli Systems and Vitalz from 1997-2001.

He entered graduated school in 2002 under the supervision of Dr. Vishwanath R. Iyer.

Hu, Z., P. J. Killion, et al. (2007). "Genetic reconstruction of a functional transcriptional regulatory network." Nat. Genet.

Killion, P. J. and V. R. Iyer (2004). "Microarray Data Visualization and Analysis with the Longhorn Array Database (LAD)." Current Protocols in Bioinformatics.

Killion, P. J., G. Sherlock, et al. (2003). "The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD)." BMC Bioinformatics **4**: 32.

Permanent address: 6800 Austin Center Blvd. #631, Austin, TX 78731

This dissertation was typed by the author.