Statistical Approaches and Rich Probabilistic Models of Biological Regulation

By

Fang Cheng

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Biology

New York University

May 2007

_____

Bhubaneswar Mishra

Acknowledgements

This dissertation would not have been possible without the help and support from many people to whom I am greatly indebted.

First of all, I thank my advisor, Bud Mishra for his support, encouragement and collaboration. It is Bud who introduced me to the computational world, and guided me into the exciting interdisciplinary field of computational biology. He has taught me the value of rigor and depth in scientific research, and has demonstrated a truly great devotion to science.

I also gratefully acknowledge my collaborators: Joan Brugge, Dennis Wall, Tobias Schmelzle, Arnaud Mailleux and Danielle Lynch at Harvard Medical School, and Eric Lightcap at Millennium Pharmaceuticals, Inc. for the intellectual swashbuckling and fun collaboration that I truly enjoyed.

I also thank my past and current colleagues. They have been my best mentors, collaborators and friends, and have filled every day of my thesis work with care, joy and excitement. Special thanks to Yi Zhou, Matthias Heymann, Raoul-Sam Daruwala, Toto Paxia, Marco Antoniotti, Bing Sun, Ofer Gill and many more. I will cherish every day that I have spent with them.

I would like to express my gratitude to Steve Small and Fabio Piano for serving on my thesis committee and for providing valuable discussions and suggestions.

I am greatly indebted to Ilya Rosenberg for his care, support and encouragement. With his love, all the difficult moments become much more endurable.

Finally, I would like to thank my family. They have selflessly provided the best in the world to me, and I owe to them everything I am now.

Abstract

Understanding biological regulation is a critical step towards our understanding of developmental and disease processes. As tremendous progress has been made in experimental technologies for sequencing the genetic code (DNA sequencing) and for quantitatively monitoring the level of expression over a developmental process in a high-throughput fashion, we are facing a great era for using computational power to extract the information that can fill up the gaps between the genetic code and the phenotypic changes, which involves the understanding of transcriptional regulation, post-transcriptional regulation, and regulatory relationship among molecular processes.

This dissertation is dedicated to the construction of probabilistic models that can realistically capture the properties of several specific types of the biological regulation, and the development of statistical or computational approaches that can "learn" these models from the currently available experimental data. In particular: (i) a novel approach named "Mixed sample Processes Enrichment Analysis" (MixPEA) that deconvolves cell heterogeneity to identify molecular processes involved in a developmental program. An application of MixPEA in breast epithelial morphogenesis uncovered promising biological hypothesis regarding the critical processes/pathways and their contributions to this morphogenetic program. (ii) the first statistical learning based approach in

automatic hypothesis inference of alternative splicing regulation from microarray-based splicing detection data. The approach distinguishes itself from the alternative methods with the ability to learn the regulatory module and the cis-regulatory code at the same time, and to integrating multi-level regulation related information to form a rich definition of the cis-regulatory information. (iii) a computational study of the transcriptional regulatory role of transcriptional factor, P63, which suggested P63's role as a regulator of an adhesion program in epithelial cells. (iv) a systematic study on how choice of sample and control groups affects the performance of motif finding algorithm.

Table of Contents

List of Figures

List of Tables

Introduction

During the past decade, the development of the fast sequencing techniques and DNA microarray technology made it possible for the first time to efficiently trace the genetic code at the genome scale and to measure the expression of thousands of genes in parallel in a single assay. These progresses in biotechnology led to the establishment and huge growth of a new field, called functional genomics. Functional genomic researches focus on the biological mechanisms that bridge the gap between substances that carry genetic information and the characteristics observed at the phenotypic level. Although its aiming problems have no difference from the interests of the traditional research areas such as developmental genetics, cell and molecular biology, the functional genomics distinguishes itself with the ability to address these questions at genome, transcriptome, and even proteome level. To achieve these large scale analyses, one need not only the "hard" technologies (e.g. the sequencing and microarray technologies) for gathering information at the genome scale, but also the "soft" technologies (i.e. the mathematical models and computational algorithms) that can extract the "meaningful" information from the large dataset and construct a testable hypothesis regarding the biological questions of interest. This thesis, as part of the research endeavor in developing novel "soft" technologies for functional genomics, addresses several open questions in methods for realistic modeling and automatic hypothesis extraction to understanding mechanisms for biological regulation:

1. A novel statistical approach that identifies a complete set of molecular processes involved in a biological program (e.g. a developmental process). This approach, which we call MixPEA, improves upon the commonly used geneset approaches by integrating a model for sample heterogeneity into the scoring function for co-expression measurement. MixPEA demonstrated significantly better effectiveness in identification of molecular processes that are under differential regulation among the different cell populations within the same sample, which is often a common, yet critical, issue for many developmental or pathological studies.

2. A probabilistic model for the regulation of alternative splicing. This model extended previously developed Bayesian network model for studying transcriptional regulatory modules into a new architecture that fit the special characteristics of alternative splicing regulation. The resulting probabilistic model contains parameters that could be learned by applying Maximum Likelihood Estimation on available large scale splicing data from microarray based or cDNA based experiments, and the learned model could be used to predict alternative splicing events under different cellular conditions.

3. A rich probabilistic model for the representation of cis-regulatory motifs. This model defines a motif in a {sequence feature + constrain} format, which provide a unified platform to integrate higher level features, such as positional constrains and combinatorial effects, into the motif representation. This model not only

benefits ab initio motif finding and mapping of identified motifs, but also has the ability to infer unknown mechanisms involved in trans-factor and cis-regulatory code type regulation.

4. A statistical learning algorithm, which consists of multiple "learning experts" for learning regulatory modules, cis-regulatory code, and constrains for cis-regulatory code all at one iterative learning program. The iterative nature of the overall learning algorithm and the hierarchical relationship among the multiple learning experts allows one expert component to leverage the learning results from other experts for achieving better initial guess for its optimization problem.

5. A systematic analysis on how choice of sample and control groups affects the performance of motif finding algorithm.

The thesis presents the above models and computational approaches in the context of specific biological studies, which focus on transcriptional level or post-transcriptional level biological regulations, and demonstrates the effectiveness of these novel approaches by careful discussion and experimental validation of the computationally constructed biological hypotheses.

1.1 Organization

My thesis is mainly a compilation of four large self-contained pieces of work. As each

study addresses a different computational challenge arisen from studies in biological regulatory mechanisms, the specific biological problem, the detailed background and motivation for each study is incorporated into the chapter in which the novel computational methods and the performance is described in detail. **Chapter 1** deals with challenges derived from sample heterogeneity; the chapter presents the MixPEA approach and its application in a breast epithelial morphogenesis time series. **Chapter 2** moves the focus from the transcriptional regulation (Chapter 1) to the post-transcriptional regulation and specifically focus on the cis-regulatory program for alternative splicing; the chapter presents the Bayesian network, motif model, and associated statistical learning algorithm that captures the splicing-specific complexity of cis-regulation, and demonstrated the effectiveness of the model and approach in identification of developmental stage-specific alternative splicing events using splicing-junction microarray data for *Drosophila melanogaster*. **Chapter 3** is similar to Chapter 1 as they both focus on process-level analysis; however, this study focus on the comparison among several cell line samples, each of which represents a homogeneous cellular background, thus a simple category-based approach, instead of MixPEA, could satisfy the needs. This study suggested intriguing hypothesis regarding P63's regulatory bias to cell adhesion genes, and extensive experimental verification were thoroughly pursued (by Danielle Lynch and et. al. at Harvard Medical School). **Chapter 4** addresses how to use intelligent choice of control groups in enrichment based statistical approach. The empirical results from analysis on the transcriptional regulatory programs during mammary acinar *in vitro* development are discussed in details.

Chapter 1 - Deconvolving Cell Heterogeneity to Define Processes Involved in Breast Epithelial Morphogenesis

1.1 ABSTRACT

Mammary epithelial cells grown in 3D basement membrane cultures form hollow spherical structures that recapitulate numerous features of alveolar structures at the termini of ductal outgrowths in the breast. The program of development of these structures involves a tightly regulated sequence of morphogenetic events. To gain insight into the transcriptional programs and biological processes associated with in vitro acinar structure development, we monitored the expression of 44,828 gene transcripts daily for 15 days in 3D cultures of MCF-10A cells. Because cells within the structures undergo opposite developmental fates, namely survival and death, single biological processes often consisted of groups of genes with correlated profiles that were themselves anti-correlated ("structured dissociation"). Despite the importance of these heterogeneously regulated processes for normal acinar development, standard gene set enrichment approaches failed to detect them. Thus, we designed a new process-enrichment analysis approach termed Mixed Process Enrichment Analysis (MixPEA) designed to handle tissue heterogeneity and complex regulation. This approach finds structured dissociation by allowing a single geneset to contain two or more clusters of correlated expression

profiles that are themselves uncorrelated, without requiring prior knowledge of the shape of the transcriptional profiles involved in the processes. MixPEA enhanced the means to detect subtle processes that are under differential regulation in the acinus or hidden by more dominant expression patterns. MixPEA also generated several biological hypotheses, including a previously unexpected role for epidermal transdifferentiation in mammary epithelial development, a switch from canonical to noncanonical WNT signaling, and a possibly pivotal role for the CYLD protein in inducing apoptosis required for lumen formation. The MixPEA method is readily scalable to more complex and heterogeneous expression samples where little is known about the molecular mechanisms at work, including developmental events and cancer.

1.2 INTRODUCTION

The mammary gland is a structurally dynamic organ, which undergoes morphogenetic changes during embryogenesis, puberty and pregnancy. Little is known about the molecular processes associated with many aspects of mammary gland development in vivo. Recently three dimensional (3D) cell culture models have been developed that are suitable for investigating certain aspects of mammary morphogenesis in vitro. These cultures promote the development of breast epithelial cells into structures that resemble certain features of the organization of luminal epithelial cells in the breast [1,2]. In these in vitro models, nonmalignant human mammary epithelial cells cultured in reconstituted basement membrane proteins undergo a series of distinct morphological changes and form spherical acini-like structures that are composed of a single layer of epithelial

cells surrounding a hollow lumen (Figure 1). This in vitro developmental process typically takes 12-15 days. Initially cells proliferate to form a solid cell cluster. After four-five days, the outer, matrix attached cells develop an axis of apical-basal polarity. This creates a dichotomy between the outer and inner cells in which the inner cells cease deposition of extracellular matrix proteins and fail to transduce signals that activate critical pathways required for cell viability. Subsequently, the inner cells display two features of stress: (1) autophagy, a process that breaks down cellular materials to provide energy under conditions of nutrient starvation [3], and (2) apoptotic cell death [4]. The latter process leads to the formation of a hollow lumen and ultimately a mature structure referred to herein as an acinus (Figure 1A) [5]. Thus, there is significant heterogeneity within cells in each structure and multiple parallel ongoing processes taking place during morphogenesis.

Although these 3D culture models do not precisely replicate the in vivo microenvironment (e.g. where acini are surrounded by a myoepithelial cell layer and a more complex microenvironment), a recent study has demonstrated that events associated with lumen formation in this model mimic many aspects associated with the clearing of excess proliferating cells in the lumen of the developing mammary gland during puberty [6]. In addition, in vitro models offer advantages for execution of molecular mechanistic studies of processes that regulate cell polarity, proliferation, and other aspects of morphogenesis, and can be used to model early non-invasive forms of breast cancer, such as carcinoma-in-situ (CIS), because these lesions remain encased by a basement

membrane barrier, and because cell re-population of the luminal space and the formation of solid cell masses are commonly observed in CIS [7,8].

A recent study analyzed the transcriptional output from human mammary epithelial cells grown in three-dimensional laminin-rich extracellular matrix on days 3, 5, and 7 of the 15 day developmental trajectory [9]. This analysis led to the identification of several sets of genes that were significantly activated or repressed during the sampled time period. The study also revealed a signature of genes repressed late during morphogenesis that could be used to classify breast cancer patients into poor and good prognosis groups with high accuracy, providing evidence that the 3D in vitro system can provide insights that are directly relevant to human biology and cancer.

Here we sought to extend this type of analysis in order to develop a complete chronological map of the molecular events that are involved in acinar in vitro development using the 3D culture system of mammary epithelial cell development involving the immortalized breast epithelial cell line, MCF-10A[10,11]. Such a map is a crucial step towards a better understanding of mechanisms that underlie normal breast development and provide information critical to an understanding of how tumor cells escape normal controls that limit proliferation and survival of normal breast cells. To generate this map, we measured the genome-wide mRNA expression daily during the 15-day time-course of acinar development in vitro, creating a comprehensive transcriptional analysis of acinar development. We then used a novel computational strategy to

characterize the functional modules (i.e. groups of genes that work together to perform a function) under significant regulation.

Knowledge-based computational methods for extracting biological insight from large genomic databases have begun to significantly expand our understanding of transcriptional regulation and functional organization of disease[12,13]. For example, Mootha et al. [14], devised a gene set enrichment method to find processes that are systematically altered in diabetic muscle, showing that genes involved in oxidative phosphorylation are coordinately down-regulated in diabetic patients. Segal et al. [13], expanded this approach by grouping gene sets with coherent signatures of expression into higher order modules in order to identify shared and unique modules across a diversity of cancers. These knowledge-based approaches have demonstrated great advantage over gene-centric analysis of microarray data, such as clustering and identification of gene signatures, because of their better tolerance to experimental noise, higher sensitivity to higher-order transcriptional behavior, and more interpretable results.

All of these knowledge-based approaches assume that a relationship between a functional module and a certain phenotype can be identified by statistically significant co-expression of the module's member genes. However, this assumption becomes problematic when analyzing an expression sample that contains cell heterogeneity (noted by [13]), as is the case with our transcriptional sample of acinar development. In the developmental progression associated with acinar morphogenesis in vitro, the inner and

outer cells of the acinus undergo highly divergent morphological changes (polarized vs. unpolarized) as well as wholly different fates (survival vs. death) that occur at different rates and by heterogeneous employment of numerous processes. Differential regulation of genes associated with changes in cell state will result in contrasting expression profiles that could reduce or ablate the significance of a predefined biological process or module, even if it is fundamental to the developmental program (Figure 2A). Also, the population sizes of the different cell types change markedly over the developmental time series, introducing yet another level of complexity, since the intensity of expression will vary. Because of these complexities, we were prompted to design a novel approach that approximately models the cell heterogeneity inherent to acinar development and that can handle unpredictable regulation of genes within a single process.

Like the standard knowledge-based approaches, our method begins with a priori defined gene sets, but unlike other approaches, it models cell mixtures by allowing for 2 or more clusters of correlated profiles that are themselves uncorrelated, a phenomenon we denote as 'structured dissociation' (Figure 2). This simple modification distinguishes structured dissociation, which is likely to be caused by biologically relevant processes (e.g. tissue heterogeneity) from random noise and provides considerable improvements in power to recognize processes that have complex regulation of constituent genes. We applied this method to the gene expression profiles of the complete in vitro acinar developmental program, identifying a broad spectrum of biological processes that contribute to the molecular organization of mammary glands. Our results reconfirm known molecular

mechanisms and reveal several previously unknown processes and genes that are likely to play important roles in acinar development.

1.3 METHODS

1.3.1 3D mammary epithelial cell culture.

MCF-10A MECs were obtained from the American Type Culture Collection (Manassas, VA) and cultured in MatriGelTM according procedures described addressed in [15].

1.3.2 Time course microarray experiments.

Epithelial cells cultured in 3D were planted onto MatrigelTM plates on Day0 and samples were gathered every day from Day2 to Day15, covering the entire in vitro acinar developmental program. Total mRNA was isolated and hybridized onto Affymetrix gene chips U133A and U133B using standard procedures, assaying 44,828 gene transcripts. Each sample was run in triplicate to provide statistical control of experimental variance.

1.3.3 Normalization and statistical analysis on microarray data.

Background correction and normalization of raw microarray data was done with the MAS5 function of Bioconductor's Affy package [16]. Normalized intensity data was log2 transformed before statistical analysis.

Two methods were used to identify a set of genes that were under transcriptional regulation. First, Statistical Analysis of Microarray (SAM) [17] timecourse analysis was used to identify genes that showed significant transcriptional change along the time axis. Second, Bioconductor's 'Limma' package [18] was used to compare the triplicate experiments between two consecutive days in order to identify transcripts under regulation in two-day time frames. p values were adjusted for multi-hypothesis testing [19] and the false discovery rate was controlled to be <= 0.05 for selecting differentially expressed genes. We chose to apply this analysis on consecutive two days to avoid potential bias against ephemeral transcriptional regulation when using the SAM timecourse analysis(Supporting Text below Figure 3).

1.3.4 Assembly of gene sets.

We assembled a collection of 1565 annotation categories (S), including 1386 Gene Ontology [20] biological processes and 179 pathways from either KEGG [21] or GenMAPP[22]. Although we chose to use published annotation databases, the starting gene sets for the MixPEA algorithm could be defined a priori via any biologically meaningful relationship, such as shared cis-regulation, protein-protein interaction, chromosome location, etc.

1.3.5 The MixPEA algorithm (Figure. 2B).

MixPEA takes two inputs, an expression time-series dataset and a knowledge-based category dataset consisting of a priori defined gene sets, S (processes, pathways,

and/or other predefined modules). The goal is to identify every biologically meaningful S even when the following challenges exist: (i) the sample for generating the expression time-series is a mix of a known number of heterogeneous cell populations, and the sample heterogeneity is intractable; (ii) no clear expectations for the transcriptional profiles that are potentially related to the biology of interest. Given the inputs, MixPEA calculates a mixed transcriptome enrichment score (mES) for each S. A high mES indicates significant transcriptional co-regulated in at least one cell population within the sample, and a low mES indicates that the members of S are not significantly co-expressed or contain certain structure in expression profiles that is unexpected based on random sampling from the expression values of all transcript measurements. The algorithm proceeds in 4 steps, and includes an optional 5th step.

Step 1. Identification of a transcriptionally regulated subset. Because acinar development is likely to involve genes that are under post-transcriptional regulation, for each S, we removed all member genes that exhibited low variance in expression across the developmental time axis from consideration. In another words, MixPEA evaluates co-expression and structured dissociation only for transcriptionally regulated genes. To do the variance filtering, we plotted the variance of all the probed genes in ascending order to observe jumps in the derivatives along the resulting curve, and subsequently chose a cutoff of 1.2 for the variance of the log2-transformed intensity values (Figure 5). This filtering method was less restrictive than the SAM and the linear model based approach that we applied in the initial analysis for differentially expressed genes. However

since the statistical significance of MixPEA is controlled at a later step for evaluating co-expression / structured dissociation (step 4), this loose filtering has the advantage of providing more candidate genes for MixPEA analysis and enhancing the ability to identify significant co-expression among weakly regulated genes.

Step 2. Approximate modeling of sample heterogeneity. To model tissue heterogeneity precisely would require, at a minimum, explicit knowledge of a gene's transcriptional output from each cell type at each time point. Unfortunately, these parameters are difficult or impossible to estimate reliably for most developmental systems, including ours. To compensate for this, we injected a K-means clustering step into our algorithm to approximately disentangle heterogeneous regulation. The number of clusters, K, was chosen to approximate the level of structured dissociation that could be caused by the expected degree of sample heterogeneity (fs; smaller values were favored to avoid over-fitting) and was upper-bounded by the maximum expected level of diversity of the transcriptional profiles within a predefined geneset (ft). Because our sample had fs=2 (reflecting the inner and outer cell types) and ft ≈5, we set K to 2 for a reasonable balance between realistic modeling and avoiding overfitting (see further discussion in Supporting Methods).

Step 3. Calculation of mixed transcriptome enrichment score (mES). A mES is a weighted average of the Pearson correlation of gene expression along the time axis at the cluster level within each S:

$$mES = \frac{1}{N}\sum_{i=1}^{K} ES_i \times N_i$$

Where N ~ the total number of genes within the current gene set that survived the variance filtering,

Ni ~ the number of genes within the cluster i,

K ~ the total number of clusters within the category, which is 1 or 2 in our study;

ES_i ~ the enrichment score computed for each cluster, as below:

$$ES_i = \frac{N_i \times (N_i - 1)}{2}\sum_{g_1=1}^{N_i-1}\sum_{g_2=g_1+1}^{N_i} Corr(g_1, g_2)$$

Where Corr(g1,g2) is the Pearson correlation between the transcription values between gene g1 and g2.

Step 4. Estimation of statistical significance of a MixPEA score. We assessed the statistical significance of MixPEA by calculating an empirical false discovery rate (FDR[23]). We estimated the probability that a category with the same number of genes that yielded the same or higher mES represented a false positive. For each S of size N gene expression vectors, we constructed a null distribution of mixed expression scores (mES*) by randomly sampling N gene expression vectors from the total expression dataset 1000 times and running these expression vectors through steps 1-3 outlined above. A quadratic fit of the top 1% of the null distribution was used to calculate an empirical FDR of 0.01 (Fig. 1D). p value cutoffs were then dynamically selected for the total size of the category in each mES score.

Step 5. Construction of significant gene set subgraphs (optional). The gene sets (S) from Gene Ontology (GO)) could be hierarchically nested in such a way as to cause redundancy, especially in cases where a general biological process was found to be regulated together with many of its sub-processes (i.e. children nodes in the hierarchy). To account for this redundancy, we built subgraphs of the biological process gene sets using a simple algorithm that begins with the biological process ontology from GO and the set of S's that produced a significant mES (W). The algorithm examines the immediate parent and children nodes of a Si randomly selected from W. If these nodes were in W, they were removed from W and joined with Si into a growing subgraph. These immediate neighbor nodes then became new Si starting points for an additional 1-hop expansion, and the algorithm proceeded until the largest possible subgraph was built. Expansion ceased if no neighbor nodes < 3 hops from any Si were found in W. We provide a more rigorous description of the algorithm in Box 1.

Below, we give a rigorous description of the algorithm:

```
Inputs:
        A set of MixPEA test identified GO categories: {SigTerms}
        Gene Ontology hierarchy (i.e. the relationship among GO terms)
Initialization:
        for each SigTerm_i ϵ {SigTerms}
                set the merging state: MergeState_i =  0;
Main algorithm:
        while  {SigTerm} ≠ Φ
                add SigTerm_1 into {CurrentSubgraph}
                add the direct neighbors of SigTerm, including the parents nodes and children nodes,
into {CurrentNeighbors}
                while {CurrentNeighbors} ≠ Φ
                        start with CNterm_1, the first element in {CurrentNeighbors}
                        if  CNterm_i ϵ {SigTerms}
                            add CNterm_i into {CurrentSubgraph}
                            delete CNterm_i from {CurrentNeighbors}
                            add the direct neighbors of CNterm_i into {CurrentNeighbors}
                        else
                            search all direct neighbors of CNterm_i
```

```
                                        if a CNterm_i's child node X ϵ {SigTerms}
                                                add both X and CNterm_i into {CurrentSubGraph}
                                                add X's direct neighbors into {CurrentNeighbors}
                                        if a CNterm_i's parent node ϵ {SigTerms}
                                                add both X and CNterm_i into {CurrentSubGraph}
only if [# of genes in X] <= 1.1*[# of genes in CNterm_i]
                                        delete CNterm_i from {CurrentNeighbor}
                                Sort {CurrentNeighbors} in increasing order of category size
                        Output {CurrentSubgraph}
                        Delete elements in {CurrentSubgraph}
```

1.4 RESULTS

1.4.1 Transcriptional assay of acinar development

To analyze changes in gene expression during morphogenesis, we isolated mRNA from triplicate samples on days 2-12 and d15 after plating MCF-10A cells in reconstituted basement membrane cultures. Samples were fed each day by replacing 25% of the medium with fresh medium in attempts to reduce feeding effects during this long period of analysis. We analyzed the initial dataset with two approaches: a time series analysis over the 15-day time-frame with the SAM (Statistical Analysis of Microarrays) package, and a short-frame comparison between consecutive two days using Limma package in Bioconductor [18]. The union of the results from these two approaches revealed 2973 genes, nearly 18% of the genes sampled by our arrays, were under significant regulation. Four major clusters could be distinguished in the global analysis of the 15-day period (Figure 4).

The most dominant transcriptional transition occurred between day 4 and day 5. More than half of the transcripts (1705) underwent significant regulation during this timeframe (depicted by two anti-correlated clusters I and III in Figure 4), 23% of which were directly or indirectly related to cell-cycle regulation and 68% of which were involved in primary metabolism, based on their association with the corresponding gene ontology terms. This strongly suggested that the initiation of growth arrest occurs at the Day4-Day5 transition in MCF-10A cells, earlier than previously reported based on detection of Ki67, a well characterized marker of cycling cells [10,11]. FACS analysis (Figure 6), phosphoRb immunostaining (data not shown, and immunoblotting of proliferating cell nuclear antigen (PCNA) confirmed that proliferation arrest occurs in the 3-6 day time frame.

Genes in Cluster IV increased in expression at later stages of morphogenesis and were maintained at high levels throughout the time course. These genes may regulate events associated with maturation of the acini. Genes in Cluster II displayed greater variability relative to each other, yet showed an overall pattern of transient upregulation during intermediate stages of development. This intermediate period corresponds to the time when the outer cells develop an axis of apico-basal polarity and a dichotomy develops between the inner and outer cells due to differential activation of intracellular signaling proteins like Akt. Thus, the timing of major transcriptional changes correlates with detectable morphological changes. This mRNA expression to phenotype correspondence together with the large percentage of genes found to be under transcriptional regulation

implied that acinar development must be regulated, at least in part, at the transcriptional level. However, to generate a complete description of the processes under transcriptional regulation it was necessary to address two challenges inherent to our transcriptional data.

First, because the acini are composed of two cell populations that could not be separated experimentally, the transcriptional assay from each day contains a sum of contributions from each cell population. The analysis is further complicated by the fact that the inner cells of the acinus die over time, while the outer cell population remains at a constant size. Thus, the intensity of expression of genes between the two cell populations varies over time, making it difficult to find processes that are differentially regulated between the two cell populations. For example, it is likely that the greater variance observed among Cluster II genes or along the time axis for a single Cluster II gene is an effect of summing at least two different transcriptional profiles. Any computational approach that does not adequately address such sample heterogeneity would miss critical functional programs that are under differential regulation in different subpopulations.

A second challenge stems from the dominance of the large coordinated set of genes that are regulated during proliferation arrest. Using standard approaches to gene set enrichment analysis, developmental processes that are under moderate transcriptional regulation and are synchronic with proliferation are likely to be missed due to the dominance of the large proliferation arrest profile. This is especially true for approaches that focus on the leading edge of ranked lists of differentially expressed genes (e.g.

[24]), as these edges are likely to include only those genes involved in the transition from a proliferative state to a more organized growth arrested state (Day4-Day5). Previous attempts to deal with tissue heterogeneity have introduced additional control groups into the experimental design [25]. Here we attempt to use a strictly computational approach to overcome the challenges described above. To do so, we devised a new analytical strategy that is able to distinguish structurally dissociated transcriptional profiles within annotated gene sets from those that are randomly dissociated, referred to as MixPEA (Figure 2; Methods). This approach relaxes the measure of co-regulation by allowing a single geneset to contain two or more clusters of correlated expression profiles (structured dissociation) and uses resampling-based approach to control the statistical significance of the structured dissociations being unexpected for randomly grouped genesets.

1.4.2 Process enrichment analysis by MixPEA

We applied MixPEA to the time series microarray dataset of acinar development using 1565 a priori-defined gene sets compiled from Gene Ontology, KEGG, and GenMapp. Our analysis identified 176 with significant enrichment scores (mES) based on comparison with the distributions of scores from size-corrected randomly generated gene sets using a false discovery rate of 0.01 (Figure 7A&B; Supplementary Table S1). These sets are comprised of 146 GO biological processes and 30 KEGG and GenMAPP pathways that are under significant regulation during acinar development.

To evaluate the performance of MixPEA, we compared the results from MixPEA

with those generated by GSEA [24]. All 72 processes found to be enriched by GSEA were also identified by MixPEA, indicating that the two approaches are consistent. However, when we ranked the gene sets identified by each approach in order of statistical significance, we discovered that the ordering was substantially different (Table 1). Many of the gene sets identified only by MixPEA had a higher rank than those identified by both approaches, suggesting that these MixPEA-only gene sets were less likely to be false positives. This improvement was a direct result of the fact that the scoring function in MixPEA combines a measure of coexpression with a model of cell heterogeneity. Thus, gene sets found only by MixPEA were the ones that contained structured dissociation rather than lower levels of coexpression.

We next grouped the enriched GO gene sets into subgraphs (Figure 12) to determine the relationships among them and to discern the extent to which they should be grouped into modules of correlated regulatory signatures. In other words, we grouped GO terms using our subgraphing procedure (see MixPEA step 5 in Methods), and then checked for correlation among the profiles across the gene sets within a subgraph. If a correlation was found, we reasoned that the gene sets should be considered as a single, globally regulated functional module. If not, we reasoned that the gene sets within a subgraph should be treated as separately regulated and independent processes regardless of their relationship structure in GO. From this analysis, we found that the 146 gene sets (S's) identified by MixPEA group into eight subgraphs representing general processes – metabolism, cell cycle, signal transduction, apoptosis, cell adhesion, response to stress, transport, and

development (Figure 5C and Figure 12 i-viii).

The eight subgraphs show distinct tree architectures that represent different transcriptional coordination. At the two extremes were the Development and Proliferation subgraphs which displayed either an uncoordinated, flat tree structure or a highly correlated nested structure, respectively. Of the 113 total developmental gene sets in our starting collection of gene sets, only 6 (5.3%) were significantly enriched. Five of these – Muscle Development, Nervous System Development, Epidermis Development, Morphogenesis, and Cell differentiation – were at similar hierarchical levels in the biological process ontology and consequently formed a flat tree structure (Figure 12 v). These development processes did not share regulatory signatures and possessed virtually no overlap of member genes, indicating that they were each associated with distinct aspects of acinar development.

In contrast, a much larger fraction of cell cycle and primary metabolism processes were identified as enriched. Specifically, 25 out 78 (32%) cell cycle categories, and 67 out of 218 (30%) primary metabolism categories had significant mES. These 92 subprocesses of cell cycle and metabolism form two large and hierarchically-nested tree structures (Figure 11 B &C). The gene sets within each tree had correlated expression patterns and shared many genes, indicating that the metabolism and cell cycle subgraphs are each regulated as a whole during the acinar development. Moreover, the metabolism and cell cycle processes were found to have broadly coherent signatures of expression

suggesting that these two general processes are regulated together as a single functional module during acinar development (Figure 7D) (average Pearson correlation = 0.813). The biggest regulatory shift among the correlated profiles of this module directly corresponded to the time at which the cells of the acinus shifted from a proliferating state to an organized, proliferation-arrested state. For this reason, we refer to this module hereafter as the "proliferation-associated component" (PAC).

1.4.3 Detecting heterogeneous regulation

It is noteworthy that 68 of the 72 gene sets identified as enriched by both our approach and GSEA were part of the PAC (Table 1). This suggested that although both approaches are equally effective at identifying the dominant program, only MixPEA could find the biological processes masked by the two challenges presented above, of which there was a total of 104. We suspected that the main reason why GSEA (or other knowledge-based gene set enrichment approaches) would miss a meaningful biological process is the insensitivity to cell heterogeneity and thus structured dissociation.

Confirming our conjecture, 82 of the 104 MixPEA-only gene sets (about 78.8%) contained structured dissociation, while only 40.3% (29 out of 72) of GSEA identified processes did (Table S1 column 3). Also, 62 of 104 MixPEA-only gene sets were not part of the PAC, i.e. not functionally directly related to the transition from a proliferating state to a growth arrested state among the cells of the acinus. Gene sets within this proliferation-unrelated component (PUC) were most likely associated with other

biologically important events that occur during acinar morphogenesis. Indeed, the PUC included, among others, genesets associated with apoptosis, cell-cell adhesion, cell-matrix adhesion, and response to oxidative stress, all of which have confirmed roles in acinar development or in related developmental processes (Table 2).

Within the Apoptosis gene set, two profiles were distinguished, one increasing and one decreasing during morphogenesis (Figure 8A). Interestingly, each profile contained genes that code for proteins with either anti- or pro-apoptotic activity, which would be predicted if two cell populations had distinct survival outcomes because cells destined for death would likely express increased pro-apoptotic proteins and decreased anti-apoptotic proteins and vice versa for cells that survive. Likewise, the Cell-Matrix gene set contains one profile of genes (including integrins α2, β5,& β6) that decrease over time during morphogenesis and another (including integrins β4and β7) that increases over time (Figure 8B). These disparate profiles imply that morphogenesis involves significant changes in cell attachment to extracellular matrix proteins. All of the significantly regulated genes in the Cell Adhesion gene set displayed an increase in expression over time (Figure 8C). These included several protocadherins (10, 14, 7, 9, B14 and B10), suggesting potential functions for these non-classical cadherin genes during morphogenesis. The Oxidative Stress gene set was also found to be significantly regulated during morphogenesis (Figure 8D). Multiple genes that are known to be induced following oxidative stress increased during morphogenesis. To address whether oxidative stress is associated with acinar cells, we probed structures for markers of

oxidative stress, and the immunoblotting results showed center cell-specific expression of oxidative stress markers (data not shown). Thus, the MixPEA revealed a previously unrecognized biological process associated with morphogenesis.

Two other processes in the PUC had less obvious association with acinar development, keratinization and epidermis development. These were both significantly up-regulated after Day7. We have previously found that there is an upregulation of markers of epidermal development in the matrix-deprived core of squamous cells that differentiate in the center of 3D structures of primary human mammary epithelial cells (HMECs) cultured in Matrigel [6]. The HMECs structures are distinguished from MCF-10A cells by the squamous differentiation, rather than apoptotic clearing, of inner cells. The MixPEA identification of regulation of Keratinization and Epidermis Development gene sets suggested that the same epidermis developmental program might also occur during in vitro acinar morphogenesis of MCF10A cells. We speculate that this differential upregulation is due to the loss of matrix protein attachment of these cells because we found that these same genes were transcriptionally upregulated when MCF-10A cells were detached from matrix. Let's do include the transpcriptional profile.

Thus, evaluation of MixPEA-only PUC gene sets confirmed that processes unrelated to the dominant phenotypic shift from proliferation to growth arrest (Day4-Day5) are involved in the acinar development. Importantly, none of these gene sets were identified by gene set enrichment approaches that do not explicitly account for cell heterogeneity.

We next asked which of the 8 subgraphs contained the largest number of processes with structured dissociation by comparing the enrichment results from the complete MixPEA pipeline with the results from an abbreviated version in which the k-means hierarchical clustering step was omitted. Compellingly, the number of enriched signal transduction processes increased by 66.7% in the MixPEA results set, while the number of enriched cell cycle categories increased only marginally (Figure 7E). (Note that the abbreviated algorithm identified an exact subset of the results from the complete MixPEA algorithm (Supplemental Table S1)). We manually examined all of the signal transduction processes identified by the complete MixPEA algorithm. We specifically focused on pathways whose regulators and targets were under transcriptional regulation. In this closer examination two pathways showed potential relevance to morphogenesis: "Wnt signaling", in which multiple receptor and co-receptors were under strong regulation during the time series, and "I-κB kinase/NF-κB cascade", in which several signaling molecules in the TNF-superfamily were under significant transcriptional regulation.

1.4.4 Differential regulation of distinct Wnt pathways

The genes within the Wnt signaling gene set fell into two groups (Figure 9A). Group I consisted of DKK1 (a well-recognized canonical Wnt pathway-specific antagonist), CTNNBIP1, an inhibitor of beta-catenin-mediated transcription (beta-catenin is downstream factor in the canonical Wnt pathway), and Frizzled 8 (a Wnt signal receptor showing closest protein similarity to the only known non-canonical Wnt receptors

(Figure 9B)). Group II consisted of canonical Wnt pathway genes, including co-receptors LRP5 and one potential non-canonical Wnt pathway inhibitor, secreted Frizzled related protein 1 (SFRP1)[26]. Group I genes showed continuous down-regulation from Day5 to Day8 and maintained lowest expression thereafter. The expression profiles of Group II genes were the exact inverse of the group I profiles, showing induction around day 8 and a continuous increase in expression through Day15. This observation suggested that a shift from high sensitivity to canonical Wnt signaling to high sensitivity to non-canonical Wnt occurred during this time period. Wnts themselves were not significantly transcriptionally regulated. This is not surprising as MCF-10A cells are not responsive to hormones (estrogen and progesterone) that are required for expression of Wnt molecules in breast tissue [27-29], Wnt would not be transcriptionally regulated in this model.

From these data it is tempting to speculate that acinar development in vivo may involve a shift from canonical to noncanonical Wnt signaling. The canonical pathway regulates cell proliferation and is essential for early mammary development[30-32], while the noncanonical pathway is known to regulate remodeling of tissues involving intercalation of cell layers. A recent study showed that a shift from canonical to noncanonical Wnt signaling is important, and possibly critical, for normal renal development [33]. In this study, it was demonstrated that when the shift was interrupted by depletion of the gene Inversin, renal cysts occurred. Given that other work has linked dysregulation of Wnt signaling to pathogenesis of certain breast cancers [28,34-36], it is tempting to speculate that an interruption of this remodeling event in normal mammary gland

development could contribute to the formation of breast tumors. Our computational analysis not only suggested a model for a shift between alternative Wnt pathways and associated it with a specific time-frame during in vitro acinar development, but also specifically highlighted the importance of the regulation on Wnt receptors/coreceptors. This points to an intriguing direction for experimental studies and may lead to new insights into processes involved in normal and tumor-associated mammary epithelial programs.

1.4.5 I-κB kinase/NF-κB cascade

The I-κB kinase/NF-κB cascade has been implicated in suppression of cell death in response to apoptotic stimuli in polarized breast epithelial 3D acinar structures[37], thus raising the possibility that this pathway may be involved in survival signaling during normal morphogenesis in our MCF-10A model. A single gene, the tumor suppressor gene cylindromatosis (CYLD), was found in two structurally dissociated, and biologically related gene sets, the I-κB kinase/NF-κB pathway and the Ubiquitin-dependent protein degradation pathway. CYLD was the only gene within this second gene set that was significantly upregulated during acinar morphogenesis. Interestingly, CYLD negatively regulates I-□B kinase/NF-κB pathway through its deubiquitinating enzymatic function, whereas the other members of this gene set positively regulate ubiquitination. As such, CYLD's profile was strongly dissociated from the other members of this gene set (Figure 10A). CYLD also showed a strong positive correlation (Pearson correlation coefficient = 0.895) with several members of the TNFα superfamily which are known

regulators of the I-κB kinase/NF-κB pathway (Figure 10B) and to our knowledge has not previously been linked to normal mammary gland development (*in vitro* or *in vivo*).

Previously CYLD has been shown to regulate the outcome of TNF signaling by affecting the balance between the conflicting apoptotic and survival pathways induced by TNF ligands [38,39]. CYLD deubuiquitination of multiple positive regulators of I-κB kinase inhibits the activation of NF-κB, thus shifting the balance towards apoptosis. In our system of acinar development, the induction of CYLD and TNFSF7, 10 and 11 occurred just prior to the detection of apoptosis, raising the possibility that CYLD may be differentially upregulated in the inner cells and may play a role in the clearing of the luminal space of the acinus. We have not been able to localize CYLD in acinar structures by immunofluorescence; therefore, we indirectly addressed whether CYLD might be implicated in death of the inner cells by examining whether the expression of CYLD is affected by loss of matrix attachment because the inner cells of acini are deprived of matrix. CYLD was strongly induced following detachment of the MCF-10A cells from matrix (Figure 10C), supporting our hypothesis that CYLD may be induced within the inner acinar and contribute to cell death by blocking the anti-apoptotic effects of NFKB signaling. Further supporting this hypothesis, downregulation of CYLD was found to induce the loss of apoptosis in the invasive breast carcinoma cell line, BT549 [40].

In our transcriptional sample of acinar development in vitro, CYLD was significantly, but weakly, expressed. Had we not used our process enrichment approach to study the

acinar developmental program, this gene would almost certainly have been missed. It was only through the identification of the structurally dissociated gene sets, "I-κB kinase/NF-κB pathway" and "ubiquitin-dependent protein catabolism", that CYLD was highlighted as important. To our knowledge, this is the first suggestion that CYLD may play a role in mammary morphogenesis. If this role of CYLD can be confirmed through further experimental analysis, in particular, experiments that could confirm that CYLD is expressed only among inner cells just prior to induction of apoptosis, this would represent an important new insight in our understanding of this developmental system, and could be involved in mammary tumor development as IKKs, which are targets of CYLD, have been implicated in breast cancer[41].

1.5 Concluding Remarks

Time series microarray experiments are becoming more common, and are important for discerning the normal development of tissues that if altered can lead to various forms of cancer. Such experiments add a layer of complexity that requires novel computational approaches to transform the expression information into a mechanistic understanding of the biological program as a whole. Some of the complexity arises from the fact that developmental programs include a mix of heterogeneous cell populations with multiple sets of processes and with varying degree of spatial and temporal compartmentalization. Thus, one must seek transcriptional explanations through statistical methods that account for cellular heterogeneity, normalization of signals across time for variation in population composition, and a possible lack of global synchronization of processes over every

population. However, an arbitrarily general statistical scheme that tries to account for unrestricted variations of the picture just described exposes itself to the danger of overfitting. The approach we describe here accounts for this danger through incorporating knowledge of the biological systems, annotation categories, and an intelligent choice of null distributions.

Our method proved effective at deciphering the molecular processes involved in mammary gland development, despite the complexity of the transcriptional signal and the dynamic heterogeneity of mammary epithelial tissue. Specifically, our method was able to detect subtle, yet biologically important, molecular processes that were missed by other similar knowledge-based approaches. These subtle processes included several that are supported by previous research, and thus unlikely to be false positives, and also many processes that were not previously known to be involved in in vitro acinar development, thereby providing much needed insights into the molecular mechanisms by which normal acinar structures develop in normal breast tissue. Our method was also able to generate testable hypothesis about the differential regulation of genes and processes associated with changes in cell state. Chief among these were two hypotheses, one concerning the shift in sensitivity to alternative types of Wnt signaling, and another concerning the possibly pivotal role of CYLD in inducing apoptosis of the inner cells. Neither role had previously been implicated in normal mammary epithelial cell in vitro development, but could be by important in the mechanisms that underlie this developmental program, as well as processes associate with breast tumorigenesis.

Our method, like other knowledge-based approaches, is an important step away from single gene analysis of expression data towards systems-level analysis. We consider MixPEA an additional tool to add to the armamentarium of methodologies being developed to make these important leaps from single genes to whole biological processes. The benefit of our analytical strategy over others is in being able to handle time and cell heterogeneous samples to find processes that are under complex regulation. Although the developmental program studied here required a simple model of heterogeneity to account for the mixture of the two cell populations characteristic of the acinus, our method is readily extensible to account for more complex tissue types, such as cancerous tissues that can contain a larger number of cell populations with a heterogeneous mixture of aberrant and normal processes. Thus, we believe the success of this approach could be readily transferred to many other biological systems of similar nature that also remain recalcitrant to classical gene set enrichment analysis.

Table 1. Top 25 genesets identified by MixPEA. Gray highlighted rows are the genesets identified also by GSEA[24]. Subgraph indicates the general biological process this gene set belongs to. FDR is the false discovery rate associated with the identification of each geneset (see Methods and Materials for details) with MixPEA approach.

MixPEA-identified genesets	Subgraph	FDR
steroid biosynthesis	Metabolism	0.0009
response to DNA damage stimulus	Metabolism	0.0011
cholesterol biosynthesis	Metabolism	0.0012
endocytosis	Transport	0.0012
fatty acid biosynthesis	Metabolism	0.0012
generation of precursor metabolites & energy	Metabolism	0.0012
glycolysis	Metabolism	0.0012
amino acid biosynthesis	Metabolism	0.0013
cytokine and chemokine mediated signaling pathway	Signal Transduction	0.0014
protein ubiquitination	Metabolism	0.0014
regulation of transcription, DNA-dependent	Metabolism	0.0014
sensory perception	Sensory Perception	0.0015
mitosis	Cell Cycle	0.0017
mitotic checkpoint	Cell Cycle	0.0018
mitotic chromosome condensation	Cell Cycle	0.0018
anion transport	Transport	0.0018
mRNA processing	Metabolism	0.0020
spindle organization and biogenesis	Cell Cycle	0.0021
synaptogenesis	Development	0.0021
keratinization	Development	0.0021
signal transduction	Signal Transduction	0.0021
vesicle-mediated transport	Transport	0.0021
central nervous system development	Development	0.0021
DNA metabolism	Metabolism	0.0022

Table 2. Summary of MixPEA identified proliferation independent components of mammary gland development

MixPEA identified molecular programs	Potential association with morphogenetic events other than proliferation arrest	Examples of transcriptionally regulated genes in each program
Response to oxidative stress	Physiological events associated with centrally located cells[42]	RGS2, LOX, APOE, DUSP1, SEPP1
Keratinization and epidermis development	Differentiation events likely to be induced by loss of matrix in centrally localized cells [mailleux]	KRT1, KRT10, PPL, SPPR1A, SPPR1B, TGM1
TNF-α signaling induced apoptotic and survival signal	Processes might contribute to selective cell death among centrally located cells	TNFSF10, TRAF4, TNFSF7
Wnt receptor signaling pathway	Potential function in both proliferation and cell differentiation[43]	CTNNBIP1, DKK1, FZD8, SFRP1[26], LRP5
Cell-cell adhesion	Establishment of polarity, cell-cell communication	PCDHB5, PCDH9, PCDHB10, PCDH7, PCDHB14, CDH13, DSC2
Cell-matrix adhesion	Establishment of polarity, outer cell differentiation	ITGB4, ITGB7, ITGB5, ITGB6, ITGA2, RAPH1, FBLN5
Vesicle-mediated transport	Endocytosis for membrane protein degradation and cell secretory function (in addition to transport of metabolites which is highly associated with proliferation arrest)	SPTBN2, TSC2, VPS4A
ERBB signaling	Potential regulatory function in proliferation, secretory cell differentiation, and cell-cell interaction[44]	GLUD1[45], FN1, LOX, RGS2, SEMA3C
Apoptosis	Antiapoptotic process and pro-apoptotic process among outer and inner cells, respectively.	BMF[46], GADD45B, APOE, FOXO3A, BCL2L12
Small GTP-mediated signaling	Potential regulatory function in cell adhesion, cytoskeletal architecture and proliferation	ARL7, CHP, RRAS, AVA3
G13 Signaling	Establishment and maintenance of polarity	CALM1, ARHGDIB, PAK3
Muscle development and Muscle contraction	Outer cell trans-differentiation to myoepithelial cells	CRYAB, MYOT, SOX6

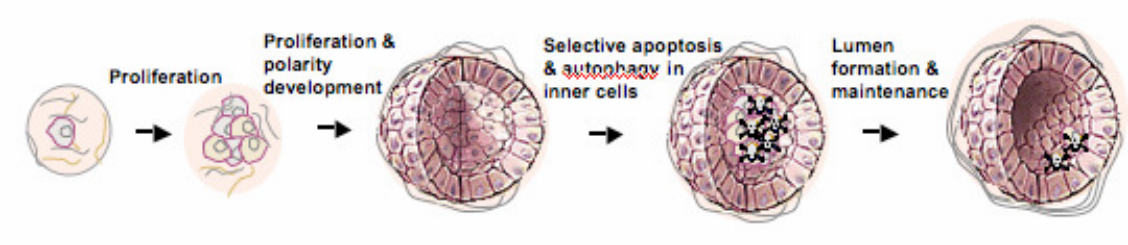


Figure 1. Human breast epithelial cell in vitro morphogenesis in 3D.

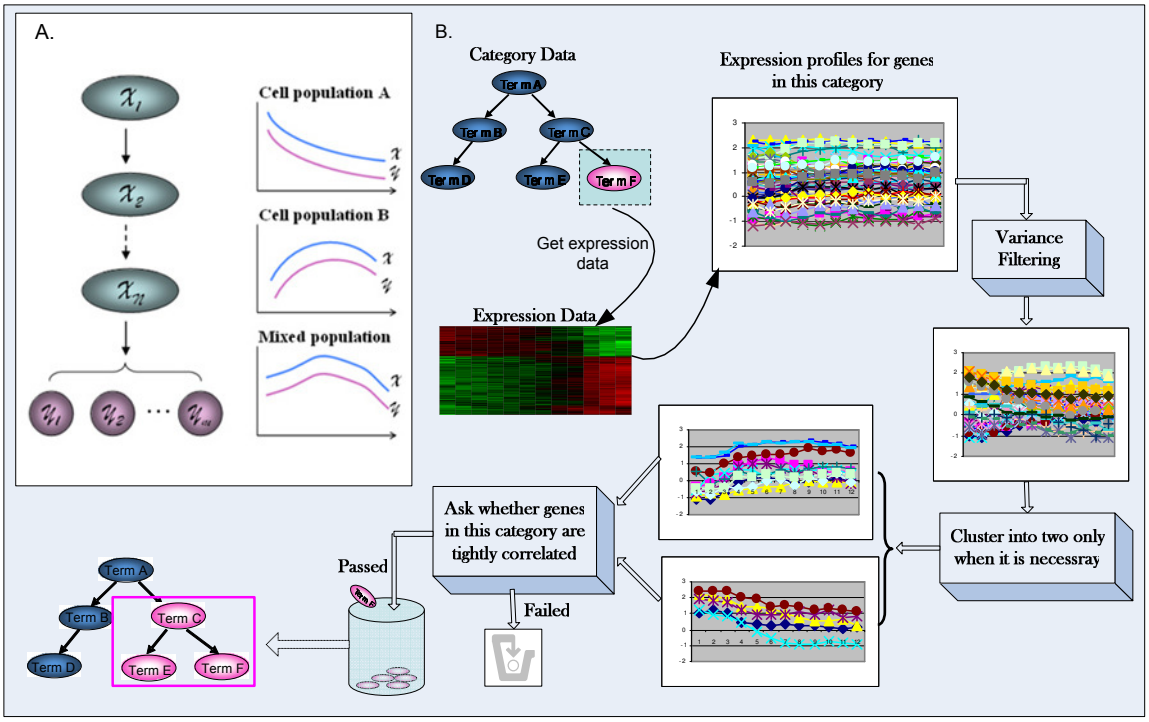


Figure 2. Summary of the MixPEA algorithm. (A) hypothetical example of a biological pathway that is under distinct transcriptional regulation in two different cell populations. The left cartoon shows the relationship among the pathway components and the plots on the right show the transcriptional profiles of each subgroup (labeled with colors and in this example, each colored curve represent the transcriptional profile of a subgroup of the pathway components (i.e. x or y)) in different cell populations (A and B), and the experimentally detected profile (the profile for the mixed population). (B) Workflow of the MixPEA approach. This computational pipeline could be divided into five major steps, which were discussed in Methods and Materials.

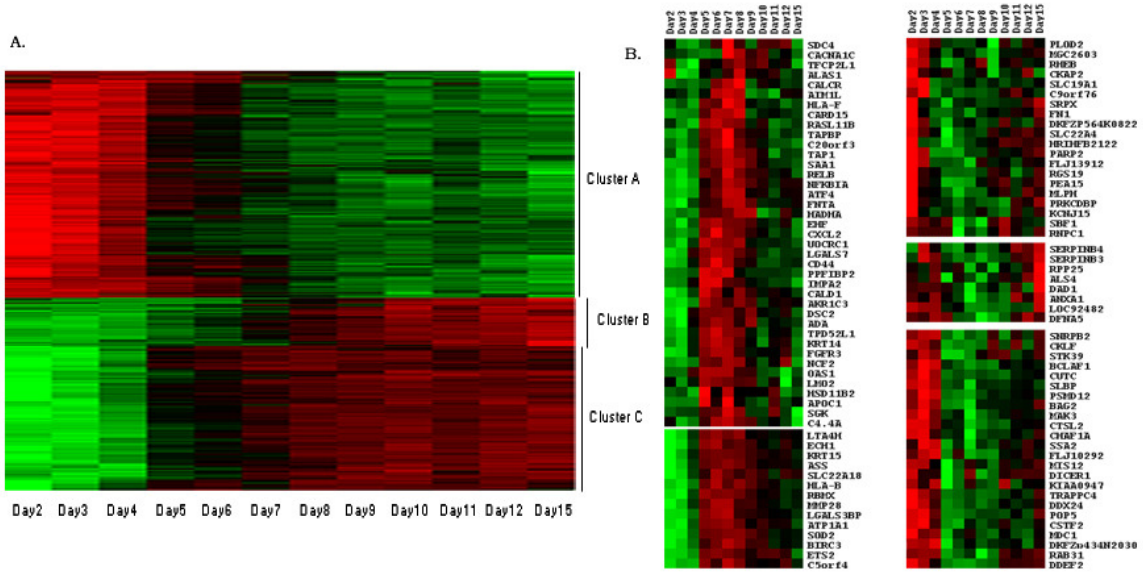


Figure 3. Major transcriptional profiles of acinar morphogenesis.Heatmap (A) shows the hierarchical clustering result for 1973 SAM[47] selected genes. Three distinguishing clusters in (A) were labeled as Cluster A, B, and C, among which Cluster A and B showed strong anti-correlation with switching of transcriptional program between Day4 to Day5 transition; Cluster B showed up regulation starting around Day8 and reached highest expression at Day15. (B) illustrates some clusters that were selected by two-day comparison method, but not SAM[47] timecourse analysis. Specifically, the left panel in B shows a down-up-down pattern and the right panel shows a up-down-up pattern, both of which might represent genes with differential expression level between inner and outer cells.

As the first step of our microarray data analysis, we filtered genes that are under transcriptional regulation during acinar morphogenesis. We applied SAM timecourse analysis, and selected 1973 probesets (1172 probesets from Affymetrix U133A and 765 probesets from Affymetrix U133B) as under significant transcriptional regulation (SAM delta = 0.55). Figure S2A presents the clustering result on SAM selected genes. We noticed that two anti-correlated transcriptional profiles (Cluter A and C in Figure S1A) strongly dominate SAM selected transcriptional regulatome. The timing of the switch on/off of transcriptional program (Day4 to Day5 transition) suggested that these clusters mainly consist of proliferation-related and growth-arrest-related genes respectively. In addition, a third cluster (Cluster B) captured a late upregulation group, in which genes are likely to contribute to lumen formation and maintenance.

However, genes involved in many biological events, such as selective cell death and cell-matrix adhesion, are likely to show differential expression between inner and outer cells. Genes exclusively expressed among inner cells would show upregulation after Day4 and downregulation after Day10 with the number of inner cells reduce during apoptosis. Missing the down-up-down profiles could be due to bias embedded in the statistical methods we used; alternatively, it could just due to the lack of regulation at transcriptional level for these events. To rule out the former possibility,

we applied a second method, which consists of a series of comparison between two continuous days. We identified 258 genes (FDR<=0.05) showed either down-up-down or up-down-up pattern, some of which were showed in Figure S2B. Thus, we took the union of the resulting sets from both methods, and use it for further analysis. Figure 1 summarized major transcriptional profiles of genes in this final set.

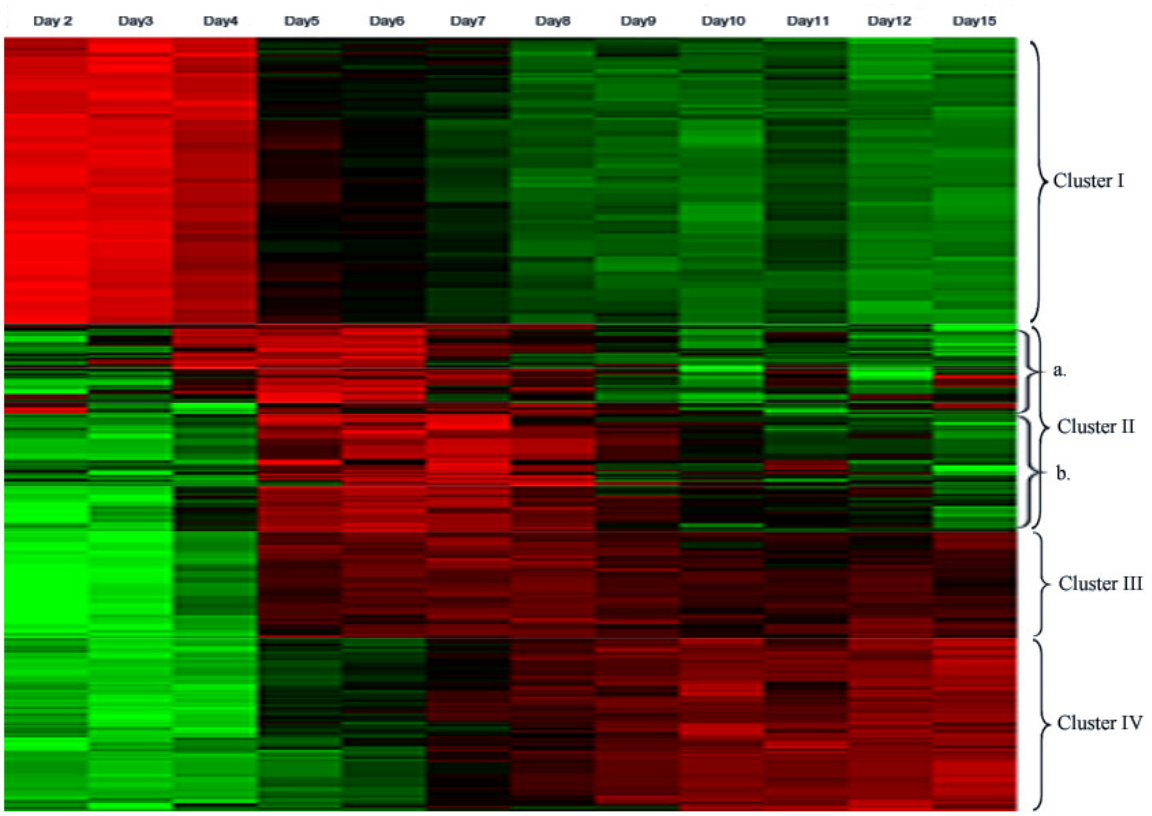


Figure 4. The global transcriptional profiles of the 15-day acinar in vitro development. Heatmap of the 2973 genes that are under significant transcriptional regulation during MFC-10A cells' in vitro development. Four major clusters are visually distinguishable in the heatmap (labels on the right). Cluster II demonstrated higher internal variety and were further divided into two subclusters labeled as a and b.

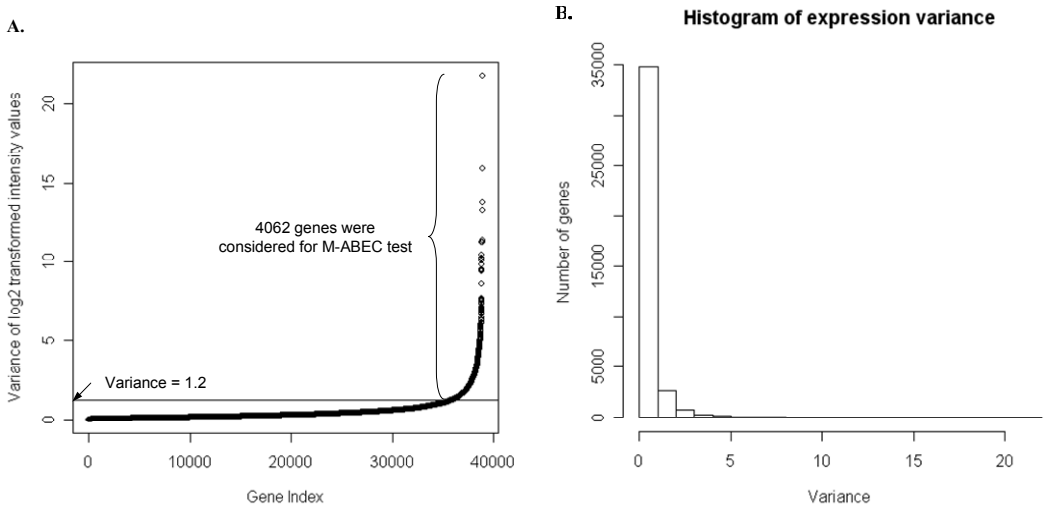


Figure 5. Choice of variance cutoff in MixPEA preprocessing. For each annotated gene, the variance was calculated for vectors representing the log2 transformed intensity values of the entire timecourse (12 time points covering Day2-Day15 mammary epithelia cell development in 3D culture). The plot (A) shows the variance values in a non-decreasing order, and the resulting curve showed a jump in the first derivative around variance 1.2, which we chose as the cutoff value for variance filtering before the MixPEA. Genes showed expression variance greater or equal to the cutoff were considered in MixPEA tests. (B) is a histogram showing the distribution of the variances of all annotated genes.

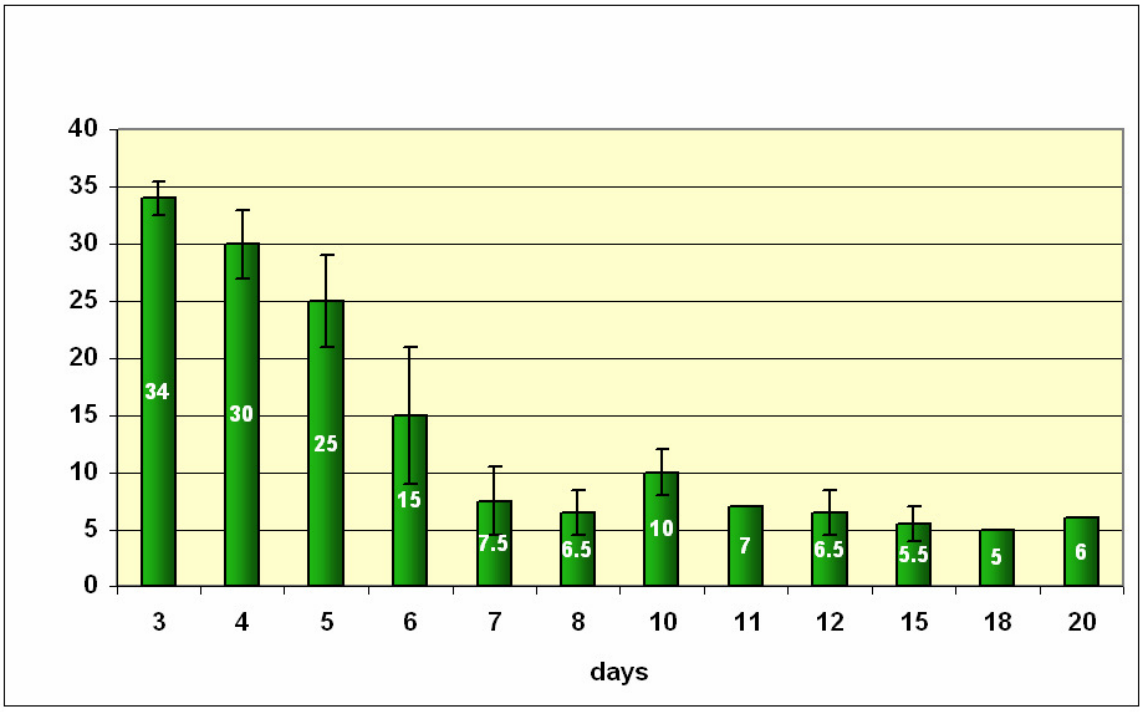


Figure 6. FACS analysis of 3D cultured MFC-10A cells.

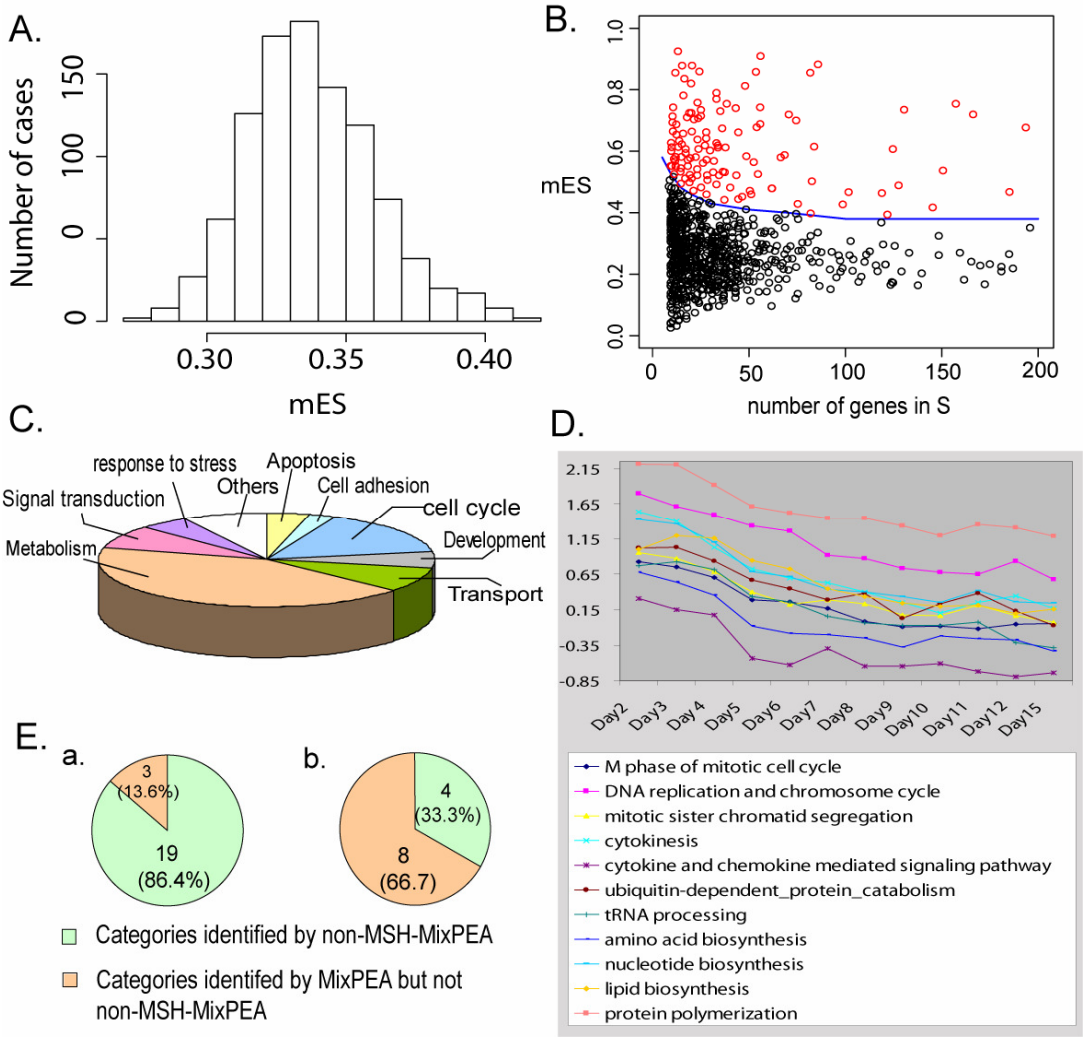
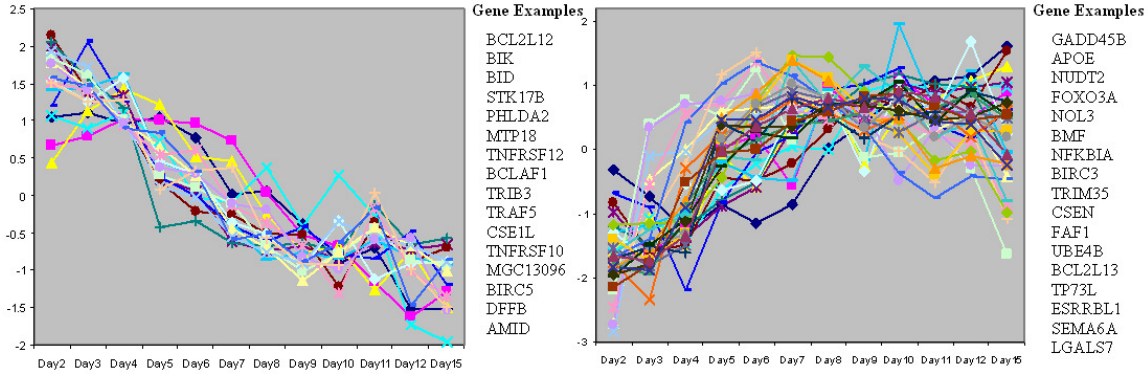


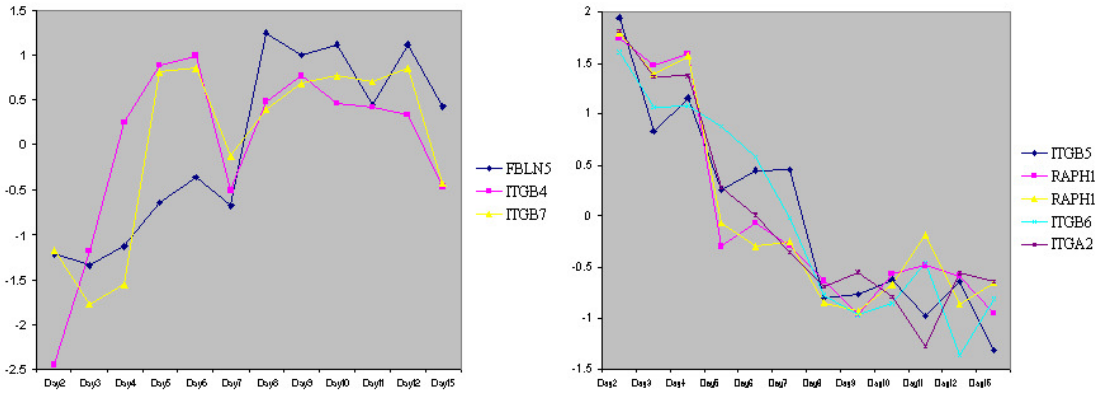
Figure 7. Summary of MixPEA identified acinar development-related molecular processes. (A) Histogram of 1000 null mES scores, each of which were calculated by running MixPEA algorithm on a randomly sampled gene set with size 70 genes. Note that the distribution shows bell shape, indicating a close to normal distribution, which verified the randomness of mES score null distribution. Other sizes of gene sets demonstrated similar normally distributed null mES scores. In (B), the red circles represent the MixPEA positive gene sets and the black circles represents the negative ones; the blue curve shows the frontier of the cutoff thresholds used for gene sets with different sizes (see Methods and Materials for details). The MixPEA positive gene sets are the ones demonstrated a significantly high mES. (C) Summary of the relative proportion of MixPEA-identified gene sets in their general biological categories. Note that in addition to the proliferation-associated processes (including cell cycle and primary metabolism) there are a significant number of biological processes were identified by MixPEA and are likely to be related to the largely unknown other aspects of the acinar development. (D) Correlation in transcriptional regulation observed among cell-cycle

processes and primary metabolism processes, suggesting the associated biological role between these processes as a proliferation-associated component in this acinar development program. (E) Comparison of the identified cell cycle gene sets (a) and signal transduction gene sets (b). Note that the non-MSH-MixPEA method (a version of MixPEA that does not model sample heterogeneity using within gene set preclustering) could identify most of the MixPEAidentified cell cycle gene sets, however, it missed a large fraction of the signal transduction processes that MixPEA could identify. This difference is likely to reflecting the difference in the extent of differential regulation on cell cycle and signal transduction processes between the inner and outer cell populations.

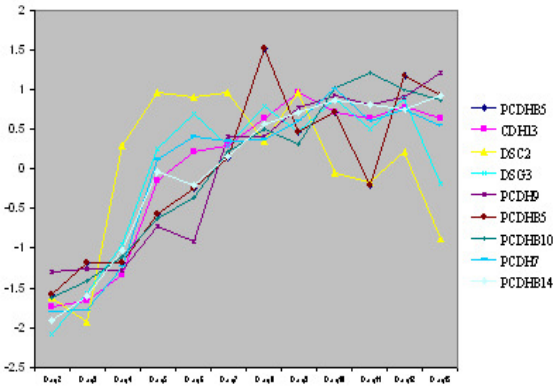
A. Regulated genes in Apoptosis genesets



B. Regulated genes in Cell-Matrix Adhesion geneset



C. Regulated genes in cell-cell adhesion genesets



D. Regulated genes in Response to Oxidative Stress geneset

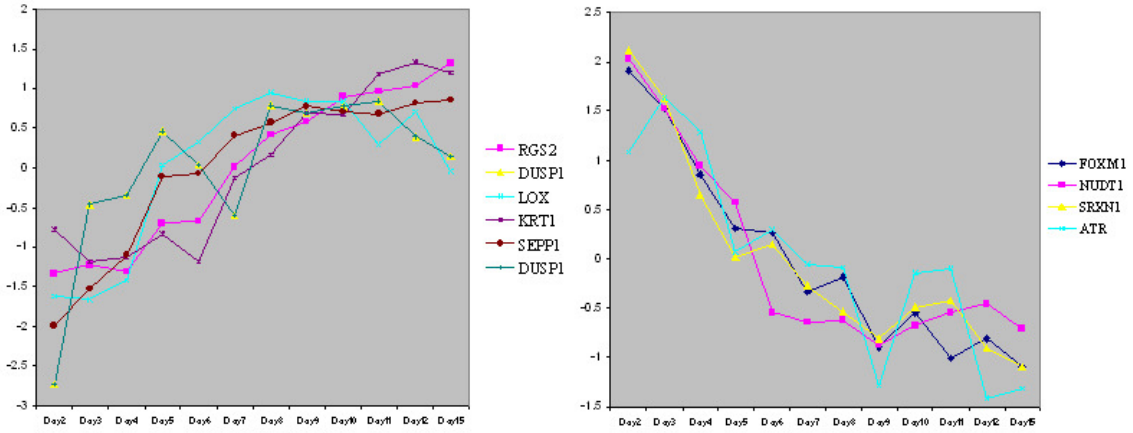


Figure 8. Examples of MixPEA-only genesets and their structured dissociation. (A) – (D) Four different MixPEA-identified genesets whose biological relevance to acinar morphogenesis were previously suggested by independent experimental studies. None of these genesets were identified by GSEA approach. Except for (C) cell-cell adhesion genesets, all the other three genesets demonstrated structured dissociation among their member genes' transcriptional profiles.

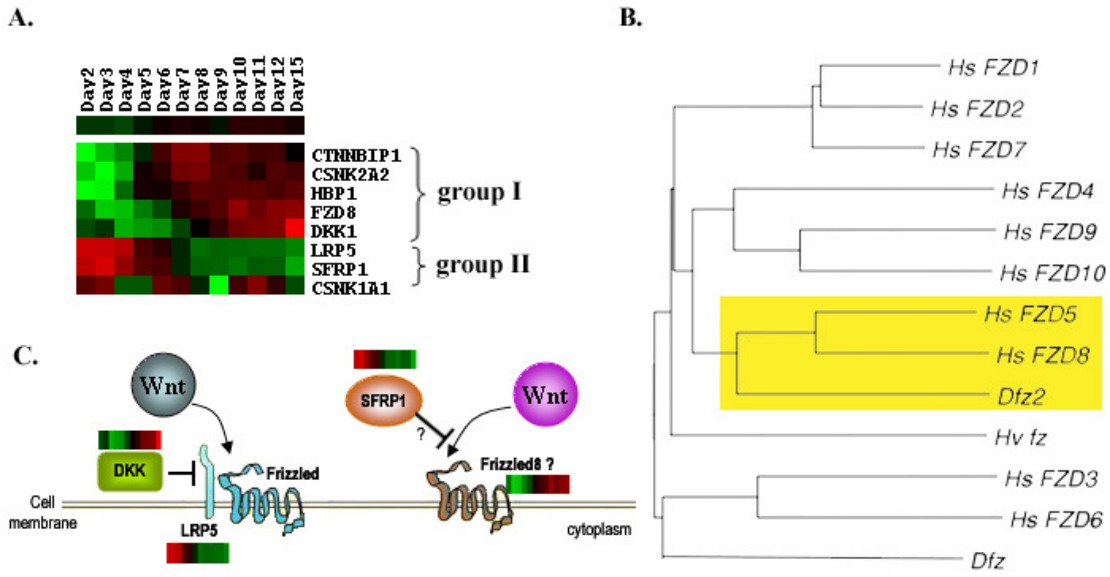


Figure 9. Hypothesis of shift in sensitivity to Wnt signaling. (A) Heatmap showing the transcriptional profile of the regulated Wnt pathway components. (B) Parsimony gene tree of the Frizzled proteins. The yellow box highlighted the branch consists of Frizzled 8, the only significantly regulated Frizzled protein in our transcriptional data, and two other Frizzled proteins (human Frizzled 5 and Drosophila Frizzled 2) that were the only two known Wnt receptors that specifically regulate noncanonical Wnt signaling. This suggested that Frizzled 8 is likely to be a Wnt receptor that is specifically functioning in the noncanonical Wnt pathway. (C) Cartoon showing a hypothetical model of the canonical (left) and noncanonical (right) signaling components at the cell membrane. The transcriptional profiles of the regulated genes were labeled with a one-row heatmap, in which the color spectrum from green to red represents the range from low transcriptional level to high transcriptional level.

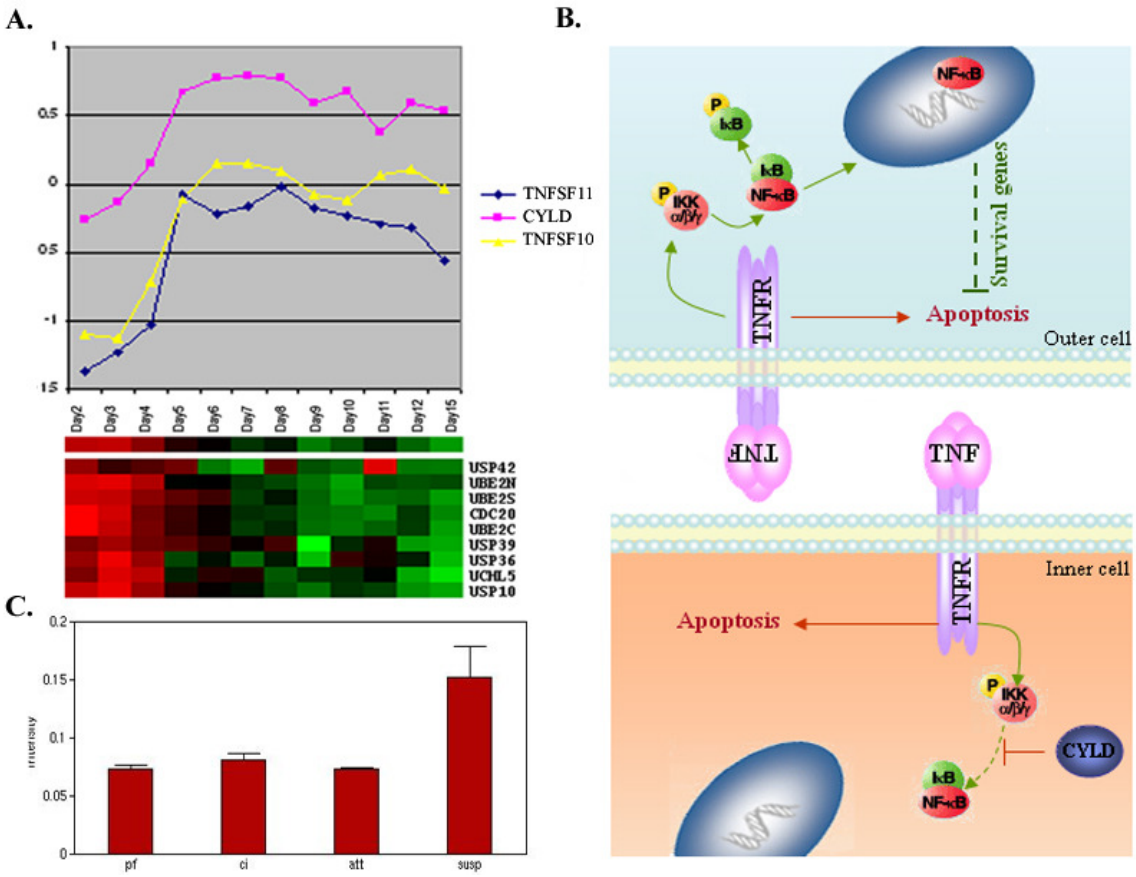
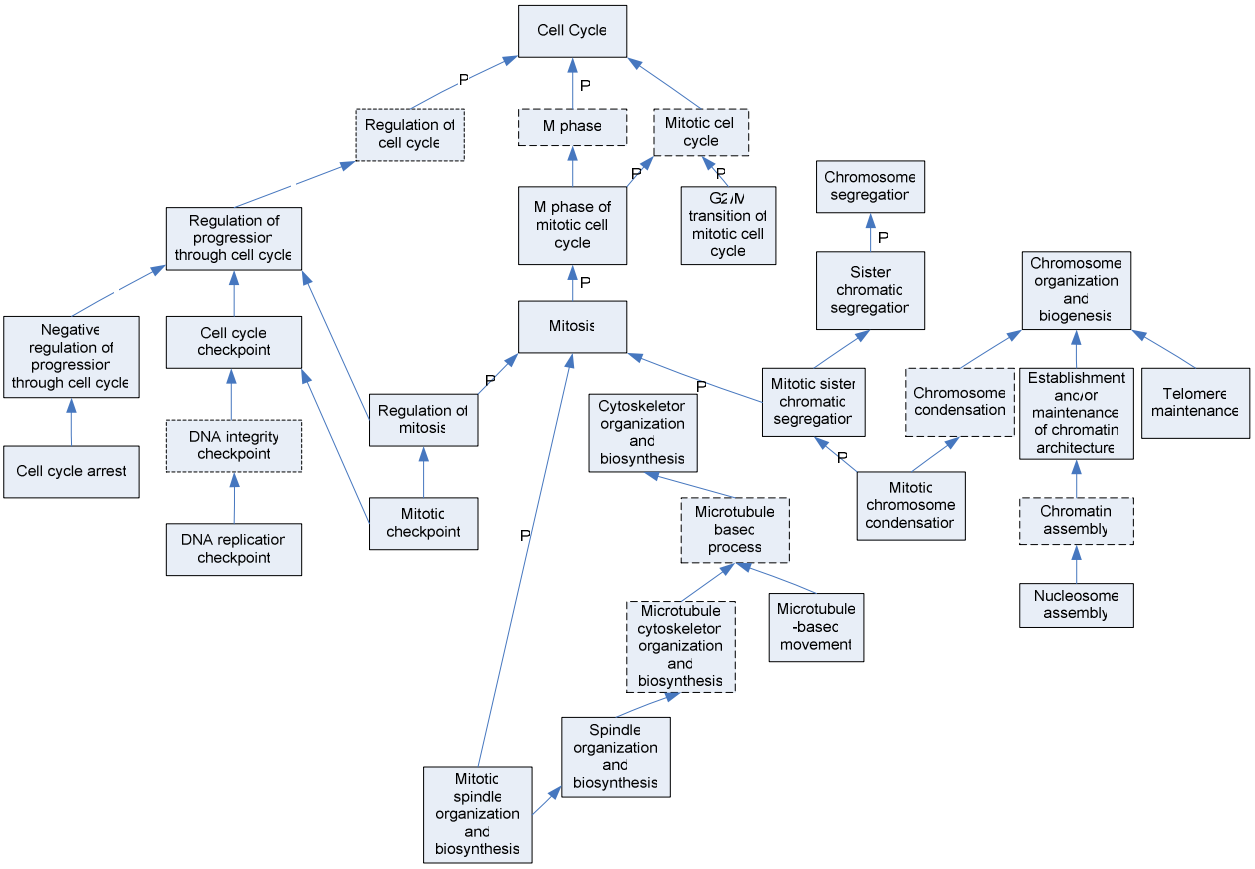


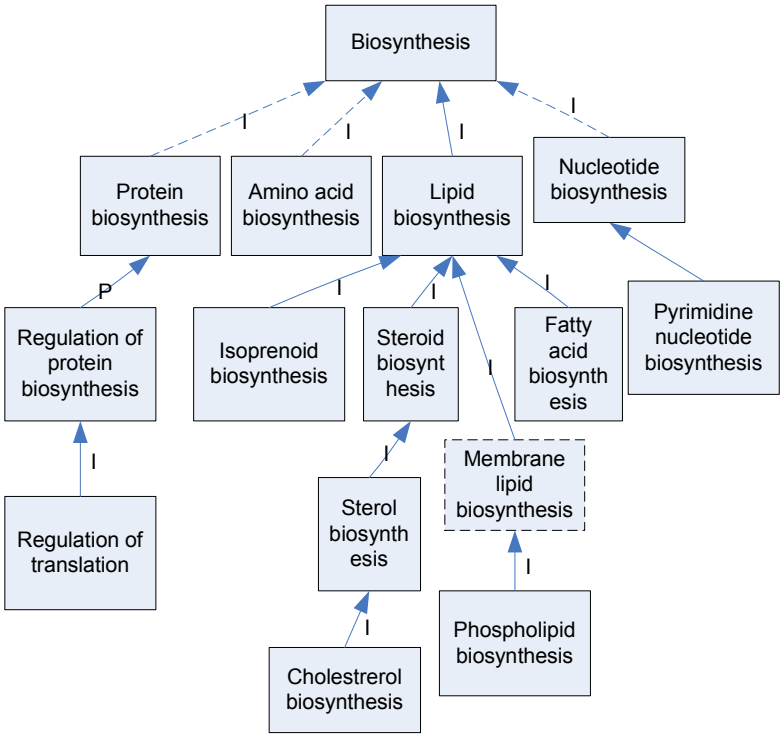
Figure 10. Model for CYLD's role in regulating differential cellular response to TNF signaling. (A) upper penal plot shows the high correlation between CYLD and TNF signaling molecules. The lower penal shows the heatmap of the transcriptional profile of the regulated member genes in "Ubiquitin-dependent catabolism" geneset, expect for the gene CYLD. Note that all these genes are positive effectors of the ubiquitinating process (i.e. having the opposite function of CYLD) and are all downregulated. In fact, CYLD was the only upregulated deubiquitinating enzyme in our timeseries. (B) is a cartoon illustrating the hypothesis of CYLD as an inner-cell specific gene that inhibits the survival signal activated by NF-κB pathway so that contributes to the inner-cell specific activation of the apoptotic process. (C) shows the hybridization signal intensity of CYLD gene under 2D culturing conditions: pf – proliferation, ci – contact inhibition, att – cell attach to plate growing condition, susp – suspension growing condition.

Figure 11. MixPEA identified eight subgraphs.

(i) Cell-cycle graph:

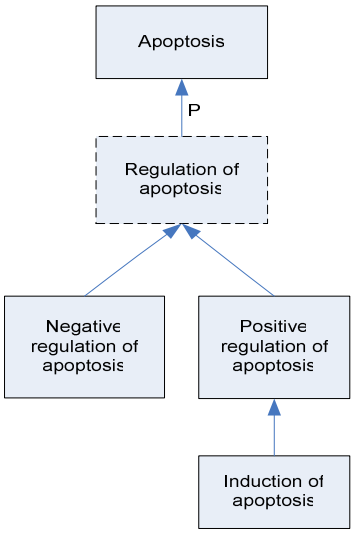


(ii) Metabolism graph:

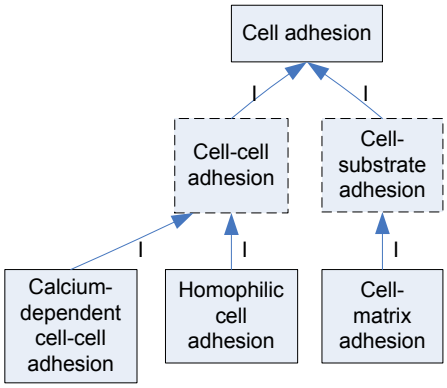


metabolism
Carbohydrate metabolism
Lipid metabolism
Protein metabolism
One carbon compound metabolism
Tricarbocylic acid cycle
Nucleobase, nucleoside, nucleotide, and nucleic acid metabolism
Amino acid metabolism
Glutamine metabolism
tRNA aminoacylation for protein translation
Generation of precursor metabolites and energy
Electron transport
Peptidoglycan metabolism
glycolysis
Fatty acid metabolism
Steroid metabolism
Membrane lipid metabolism
Sphingolipid metabolism
Protein folding
proteolysis
Ubiquitin dependent protein catabolism
Protein complex assembly
Nucleosome assembly
Protein polymerization
Protein modification
Protein amino acid dephosphorylation
Protein amino acid phosphorylation
Ubiquitin cycle
Protein amino acid ADP ribosylation
Protein ubiquitination
transcription
Regulation of transcription
Regulation of transcription, DNA dependent
Negative regulation of transcription
Negative regulation of transcription, DNA dependent
Regulation of transcription from RNA polymerase II promoter
Regulation of transcription from RNA polymerase III promoter
Negative regulation of transcription from RNA polymerase II promoter
RNA metabolism
RNA processing
rRNA processing
tRNA processing
mRNA processing
Nuclear mRNA splicing, via spliceosome
mRNA catabolism
mRNA catabolism nonsense mediated decay
DNA metabolism
DNA recombination
DNA replication
DNA replication initiation
Establishment and/or maintenance of chromatin architecture
DNA repaire
Response to DNA damage stimulus

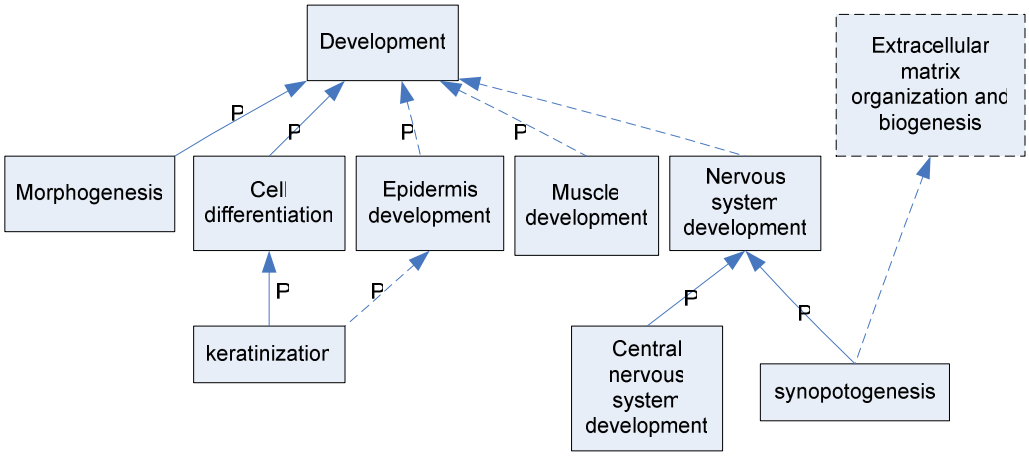
(iii) Apoptosis graph:



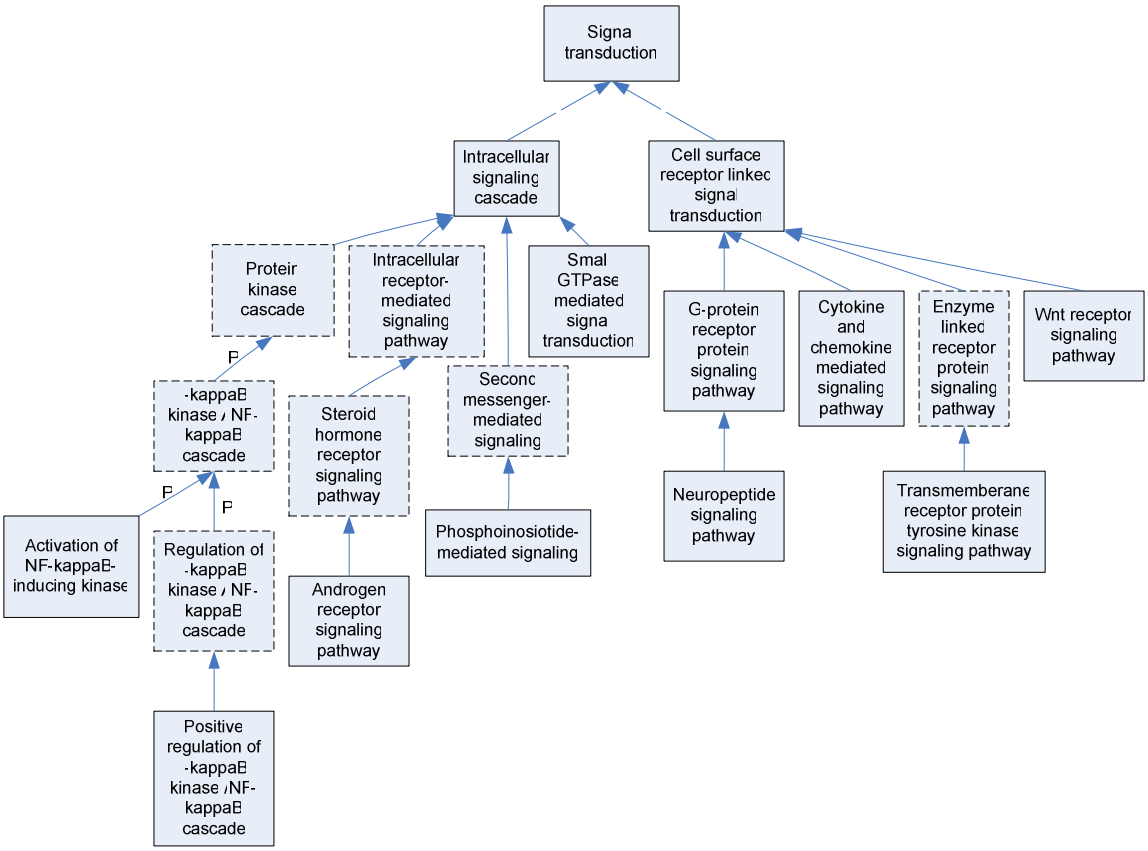
(iv) Cell adhesion graph:



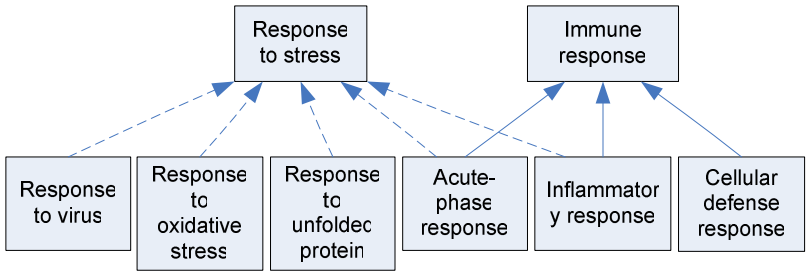
(v) Morphogenesis graph:



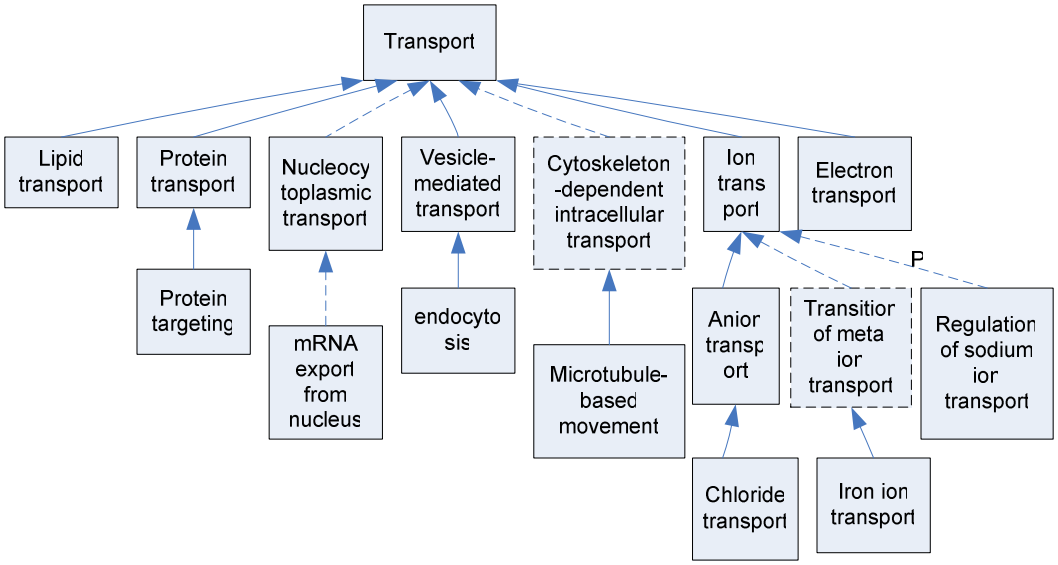
(vi) Signal transduction graph:



(vii) Response to stress graph:



(viii) Cellular transport graph:



1.6 REFERENCES

1. Schmeichel KL, Bissell MJ (2003) Modeling tissue-specific signaling and organ function in three dimensions. J Cell Sci 116: 2377-2388.

2. Petersen OW, Ronnov-Jessen L, Howlett AR, Bissell MJ (1992) Interaction with basement membrane serves to rapidly distinguish growth and differentiation pattern of normal and malignant human breast epithelial cells. Proc Natl Acad Sci U S A 89: 9064-9068.

3. Mills KR, Reginato M, Debnath J, Queenan B, Brugge JS (2004) Tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) is required for induction of autophagy during lumen formation in vitro. Proc Natl Acad Sci U S A 101: 3438-3443.

4. Debnath J, Mills KR, Collins NL, Reginato MJ, Muthuswamy SK, et al. (2002) The role of apoptosis in creating and maintaining luminal space within normal and oncogene-expressing mammary acini. Cell 111: 29-40.

5. Jayanta Debnath JSB (2005) Modelling glandular epithelial cancers in three-dimentional cultures. Nature Reviews 5: 675-688.

6. Mailleux AA, Overholtzer M, Schmelzle T, Bouillet P, Strasser A, et al. (2007) BIM Regulates Apoptosis during Mammary Ductal Morphogenesis, and Its Absence Reveals Alternative Cell Death Mechanisms. Dev Cell 12: 221-234.

7. Connolly JL, Boyages J, Schnitt SJ, Recht A, Silen W, et al. (1989) In situ carcinoma of the breast. Annu Rev Med 40: 173-180.

8. Tsikitis VL, Chung MA (2006) Biology of ductal carcinoma in situ classification

based on biologic potential. Am J Clin Oncol 29: 305-310.

9. Fournier MV, Martin KJ, Kenny PA, Xhaja K, Bosch I, et al. (2006) Gene expression signature in organized and growth-arrested mammary acini predicts good outcome in breast cancer. Cancer Res 66: 7095-7102.

10. Jayanta Debnath KRM, et. al. (2002) The role of apoptosis in creating and maintaining luminal space within normal and oncogene-expressing mammary acini. Cell 111: 29-40.

11. Debnath J, Muthuswamy SK, Brugge JS (2003) Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. Methods 30: 256-268.

12. Eran Segal NF, Daphne Koller, Aviv Regev (2004) A module map showing conditional activity of expression modules in cancer. Nature Genetics 36: 1090-1098.

13. Vamsi K Mootha CML, et. al. (2003) PGC-1alpha-responsive genes involved in xoidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics 34: 267-273.

14. Gordon K. Smyth JM, Hamish S. Scott (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics 21: 2067-2075.

15. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102: 15545-15550.

16. David A. Glesne ea (2006) Subtractive transcriptomics: establishing polarity

drives in vitro human endothelial morphogenesis. Cancer Research 66: 4030-4040.

17. Caldwell GM, Jones C, al. e (2004) The Wnt antagonist sFRP1 in colorectal tumorigenesis. Cancer Research 64: 883-888.

18. Brisken C, Heineman A, Chavarria T, al. e (2000) Essential function of Wnt-4 in mammary gland development downstream of progesterone signaling. Genes & Development 14: 650-654.

19. Gattelli A, Cirio MC, Quaglino A, Schere-Levy C, Martinez N, et al. (2004) Progression of pregnancy-dependent mouse mammary tumors after long dormancy periods. Involvement of Wnt pathway activation. Cancer Res 64: 5193-5199.

20. Inadera H, Dong HY, Matsushima K (2002) WISP-2 is a secreted protein and can be a marker of estrogen exposure in MCF-7 cells. Biochem Biophys Res Commun 294: 602-608.

21. van Genderen C, Okamura RM, Farinas I, Quo RG, Parslow TG, et al. (1994) Development of several organs that require inductive epithelial-mesenchymal interactions is impaired in LEF-1-deficient mice. Genes Dev 8: 2691-2703.

22. Andl T, Reddy ST, Gaddapara T, Millar SE (2002) WNT signals are required for the initiation of hair follicle development. Dev Cell 2: 643-653.

23. Chu EY, Hens J, Andl T, Kairo A, Yamaguchi TP, et al. (2004) Canonical WNT signaling promotes mammary placode development and is essential for initiation of mammary gland morphogenesis. Development 131: 4819-4829.

24. Simons M, Gloy J, Ganner A, Bullerkotte A, Bashkurov M, et al. (2005) Inversin, the gene product mutated in nephronophthisis type II, functions as a molecular switch

between Wnt signaling pathways. Nat Genet 37: 537-543.

25. Brown AM (2001) Wnt signaling in breast cancer: have we come full circle? Breast Cancer Res 3: 351-355.

26. Polakis P (2000) Wnt signaling and cancer. Genes & Development 14: 1837-1851.

27. Weeraratna AT, Jiang Y, al. e (2002) Wnt5a signaling directly affects cell motility and invasion of metastatic melanoma. Cancer Cell 1: 279-288.

28. Weaver VM, Lelievre S, Lakins JN, Chrenek MA, Jones JC, et al. (2002) beta4 integrin-dependent formation of polarized three-dimensional architecture confers resistance to apoptosis in normal and malignant mammary epithelium. Cancer Cell 2: 205-216.

29. Trompouki E, Hatzivassiliou E, Tsichritzis T, Farmer H, Ashworth A, et al. (2003) CYLD is a deubiquitinating enzyme that negatively regulates NF-kB activation by TNFR family members. Nature 424: 793-796.

30. Brummelkamp TR, Nijman SMB, Dirac AMG, Bernards R (2003) Loss of the cylindromatosis tumour suppressor inhibits apoptosis by activating NF-kB. Nature 424: 797-801.

31. Wang L, Baiocchi RA, Pal S, Mosialos G, Caligiuri M, et al. (2005) The BRG1- and hBRM-associated factor BAF57 induces apoptosis by stimulating expression of the cylinromatosis tumor suppressor gene. Molecular and Cellular Biology 25: 7953-7965.

32. Eddy SF, Guo S, Demicco EG, Romieu-Mourez R, Landesman-Bollag E, et al. (2005) Inducible IkappaB kinase/IkappaB kinase epsilon expression is induced by CK2 and

promotes aberrant nuclear factor-kappaB activation in breast cancer cells. Cancer Res 65: 11375-11383.

33. Lucy Erin O'Brien MMPZ, Keith E. Mostov (2002) Building epithelial architecture: insights from three-dimensional culture models. Nature Reviews Molecular Cell Biology 3: 531-537.

34. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.

35. Virginia Goss Tusher RT, Gilbert Chu (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America 98: 5116-5121.

36. Yoav Benjamini Dy (2001) The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29: 1165-1188.

37. Michael Ashburner CAB, et. al. (2000) Gene Ontology: tool for unification of biology. Nature Genetics 25: 25-29.

38. Minoru Kanehisa SG, et. al. (2002) The KEGG databases at GenomeNet. Nucleic Acids Research 30: 42-46.

39. Kam D. Dahlquist NS, et. al. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nature Genetics 31: 19-20.

40. Anat Reiner DY, Yoav Benjamini (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19: 368-375.

41. James A. Mobley RWB (2004) Estrogen receptor-mediated regulation of

oxidative stress and DNA damage in breast cancer. Carcinogenesis 25: 3-9.

42. Cadigan KM, Nusse R (1997) Wnt signaling: a common theme in animal development. Genes & Development 11: 3286-3305.

43. Iris Alroy YY (1997) The ErbB signaling network in embryogenesis and oncogenesis: signal diversification through combinatorial ligand-receptor interactions. FEBS Letters 410: 83-86.

44. Jie Feng AVR, et. al. (2004) Processing enzyme glucosidase II: proposed catalytic residues and developmental regulation during the ontogeny of the mouse mammary gland. Glycobiology 14: 909-921.

45. Schmelzle T, Mailleux AA, Overholtzer M, Carroll JS, Solimini NL, et al. (2007) Functional role and oncogene-regulated expression of the BH3-only factor Bmf in mammary epithelial anoikis and morphogenesis. Proc Natl Acad Sci U S A 104: 3787-3792.

Chapter 2 - Learning the multi-level cis-regulatory code of alternative splicing

2.1 Abstract

Alternative splicing is one of the most significant mechanisms for regulating gene expression and enhancing protein diversity, and is also a natural source of disease-causing errors. The regulation of alternative splicing involves multiple control mechanisms, which are all realized through the interplay of cis-acting sequences and trans-acting factors. Significant progress has been made in computationally identifying alternative splicing-related cis-regulatory code. However, the existing models are restricted by an oversimplified assumption that prohibits the combinatorial effects among different control mechanisms. Here, we present a novel approach, which consists of a mixture of "experts" that are specialized in learning a partition of alternatively spliced sites into regulatory modules, learning the module-specific short sequence signatures (motifs), and learning the biologically meaningful constrains associated with each motif (such as motif combination, motif spatial distribution and basal splice signals). The algorithm iterates over the experts and optimizes each expert's learning results for the simplest best overall explanation of the experimentally observed alternative splicing profiles. Using Drosophila developmental dataset generated with splicing-junction probes[48], our approach identified both previously known and unknown cis-regulatory motifs, associated their regulatory function with specific splicing profiles, and more

importantly, for some motifs, our program selected intriguing higher level properties that could be critical in controlling the specificity of the motifs' regulatory function.

2.2 Introduction

Alternative pre-mRNA splicing is a major cellular process by which functionally diverse proteins can be generated from the primary transcript of a single gene. It was suggested that at least 60% of human genes, about 55% of mouse genes, and 18% of Drosophila melanogaster genes are alternatively spliced. In the passed decade, alternative splicing has increasingly been recognized as a major regulatory process in development and has been linked to various common human diseases[49], including cancers[50] and Alzheimer's[51,52]. One major challenge in understanding alternative splicing is to understand the mechanism for its regulation.

Alternative splicing are highly regulated events. Its fidelity to the cellular contexts (e.g. tissue types and developmental stages) and specificity to the target genes are controlled by the interaction between trans-factors (usually proteins) and cis-regulatory code (a short sequence signature). Such trans-factor and cis-code based regulatory mechanism is commonly seen in biology. For example, it is the major mechanism for regulation of gene transcription. However, regulation of alternative splicing seem to have a greater level of complexity as there are more variety in ways of functioning for both trans-factors and cis-code, and the regulation can potentially involve the combinatorial effects among different types of trans / cis – signals. In the rest of the introduction section, we firstly

summarize the current biological models for alternative splicing regulation, and review the recent progress made by applying computational approaches. We will analyze the gaps between the assumptions made for computation and the biological models for the regulatory mechanism, which lead to the motivation of our work.

2.2.1 Current models of alternative splicing regulation

All genes that contain an intron region will go through a post-transcriptional process, in which the intronic regions are removed and exons are juxtaposed and ligated before being translated to proteins. Alternative splicing refers to the natural variants in splicing by linking together different 5' and 3' splice sites, which is different from the most frequent form of splicing of a gene (the constituent splicing). By regulatory mechanism of alternative splicing, we mean the mechanism involved in determining when and where an alternative splice form will be preferred to the constituent one. The traditional view is that the regulatory mechanism differentiates from the operating mechanism of splicing; the operating mechanism refers to the process of assembling basic splicing machinery (i.e. spliceosome), recruiting the spliceosome onto the 5' and 3' splice sites (namely the basal splicing signal), and the enzymatic reaction of cutting and linking together a 5' and a 3' site; the regulatory mechanism consists of a regulator (a trans factor, usually protein), which is not part of the splicosome, and a short sequence tag (the cis-regulatory sequence / motif), which sits within the pre-mRNA somewhere other than the basal splicing signals. It has been suggested that the motifs for alternative splicing include exonic splice enhancers (ESE), exonic splice suppressors (ESS), intronic splice enhancers (ISE), and

intronic splice suppressors (ISS). Although the recognition of basal splice sites by splicosome could also be considered as the trans-factor and cis-code type interaction, the traditional model of alternative splicing regulation considers it as part of the operating system, which is separated from the regulatory system, and regards the spliceosome as one ubiquitously and constantly available machinery and extra system (i.e. the regulatory system) is required for alternating it's splicing preference.

Recent progress in research on alternative splicing regulation has challenged the traditional model from multiple aspects. First, components of splicosome were suggested to be under extensive transcriptional regulation[53,54] and modification (e.g. phosphorylation) or change in relative expression level among core component of splicing factors could cause alternative splicing[55,56]; second, redundancy exists among the constituents of splicosome, so that alternative choice of splicosome components may exist over different cellular context[54,57]. These discoveries indicated that the basal splicing machinery was not consistent, and its variants could potentially lead to different splicing preference (i.e. preference to basal splice sites and preference to other regulatory trans-factors).

Moreover, emerging studies suggested that basal splicing signals (e.g. 5' ss, 3'ss, and branch site) could store information regarding the regulation of alternative splicing. As an example, according to the traditional model of alternative splicing, the variants of 5' splice sites, which yields weaker affinity to U1 small nuclear ribonucleoprotien particle

(snRNP), were equally treated as sub-optimal 5' splice sites, so that they would share similar regulatory mechanism which regulates the splice decision upon the suboptimal 5'ss by providing additional help on recruiting the basal splice machinery onto the suboptimal splice sites. However, there have been evidences suggested that the differential binding of the U1 snRNP to the 5' splice site not only affects 5' splice site recognition, but also plays a fundamental role in 3' exon selection[58], and the information regarding the 3' ss selection are likely to be stored in 5'ss in a way other than the overall U1 snRNP affinity. In another words, differential 3'ss selection decisions could be made solely based on the regulatory information stored with the 5'ss.

These conflicting evidences to the traditional model of alternative splicing suggested the necessity to relax the separation between the operating system and the regulatory system, and to extend the definition of cis-regulatory code of alternative splicing to include the special subset of basal splicing signals that can differentially affects splice consequence by sensing the states of the spliceosome or other trans-factors mediated regulation. Such relaxation on the definition of a regulatory system also post a great new level of complexity, as multiple types of cis-regulatory code could affect the splice decision at a splice sites in a combinatorial way. For example, two suboptimal 5'ss contain the same ESE in their flanking exons, when the corresponding trans-factor binds to the ESE, it will assist the recruitment of spliceosome onto both suboptimal 5'ss and "turn on" the splice at these suboptimal 5'ss; however, this would not have any effect on the 3' exon selection, and different splice decision could be made for the downstream 3'ss if one, but not the

other, contains a cis-code of another form (e.g. a weak branch site) which can contribute to 3'ss selection.

2.2.2 Accomplishments and limitations of the existing computational approaches

Computational approaches have been playing a critical role in research of alternative splicing. Based on the fundamental question of interest, these approaches could be grouped into two families: the detection approaches (i.e. the methods for computationally detect the alternative splicing events and their associated cellular context) and the approaches designed for uncovering the regulatory mechanisms that underlie the detected alternative splicing events.

Assisted by the significant progress on sampling tissue-specific EST libraries and probing splicing events with microarray-based technique, the detection approaches have been applied to generate satisfying quality splicing map over multiple genomes and diverse developmental/pathological conditions. This provided tremendous amount of opportunities for computationally studying the regulatory mechanisms of alternative splicing, which is the focus of this paper.

One major focus of the analysis on regulatory mechanism is to identify the cis-regulatory code that controls the transcript-specificity of a splicing regulator's function. A variety of computational approaches has been developed, and they share a similar

working scheme: grouping together the alternatively spliced sites that were assumed to share a similar regulatory mechanism (the positive group), constructing a control group consisting of splice sites that were not alternatively spliced or were not expected to be having the kind of regulation of interest, and analyzing the flanking exons/introns (DNA or RNA sequences) of the splice sites using usually statistics-based methods to find one or multiple short sequence features that can tell apart the positive group and control group.

The computed short sequence features (cis-code) are represented with scoring matrices (the most often used is positional weight matrix) that, for a given sequence, can be used to calculate the probability of the sequence containing the cis-regulatory code being represented. However, such representation could not fully represent a biologically meaningful cis-code since there could be other features associated with a matching sequence and is critical to the functionality of the cis-code. According to the model for alternative splicing regulation, these constrains could include positional specificity (Cpos, e.g. exonic cis-code is unlike to function, at least not the same, when it's located in the intronic ragion), the constrains on the distance to the targeting splice sites (Cdist), properties of the targeting basal splice signals (Cbss), and etc. Among this biologically meaningful constrains, Cpos is the easiest to model by separating the motif-finding process among exonic and intronic regions, and has been considered by most of the existing approaches, while there has been no approaches that can integrate the other potentially also critical constrains into the representation of a cis-code, and automatically

identify these constrains while computing the scoring matrix representation of a cis-code.

Moreover, the choice of positive and control groups is a fundamental factor to the performance of these approaches. For example, if all alternatively spliced exons were regarded as positive group and constitutive exons as control group, it is more likely to identifying common ESEs that assist spliceosome recruitment than identifying tissue-specific/development-specific exonic cis-code. Previously, the computational identification of splicing related cis-code has been mainly focusing on the former (the common ESE/ESS/ISE/ISS), because the later requires not only the annotation of an alternatively spliced sites, but also the high quality annotation regarding the condition (e.g. tissue types, developmental stages) where certain alternative splice form is chosen.

Aiming at a realistic model of the cis-regulatory code for alternative splicing regulation and a robust computational algorithm that can automatically "learn" the model from experimental data, we developed a novel approach, which consists of a mixture of "experts": the module expert optimizes the partition of alternatively spliced sites into modules for better separation of distinct regulatory mechanisms, the motif expert optimizes the matrix representation of a motif for better explanation of the partition of splice sites over modules, and the constrain experts identify biologically meaningful Cpos, Cdist, and Cbss that can be add to the representation of a learned motif in addition to the scoring matrices. The algorithm iterates over the experts and optimizes each expert's

learning results for the simplest best overall explanation of the experimentally observed alternative splicing profiles.

We applied this approach to a microarray dataset designed for probing splicing events over six developmental stages of Drosophila melanogaster, and identified both previously known and unknown cis-regulatory codes, and quantified the contribution of these cis-code or their combinations to different regulatory modules over the six-developmental stages. The identified cis-regulatory codes were represented in the {PWM + constrain} form, so that this automated "learning" process directly suggest intriguing hypotheses regarding the information related to the functionality of a motif, which could beyond the sequence feature itself (multi-level cis-regulatory code).

2.3 Methods and Materials

2.3.1 Splicing data for Drosophila development (Figure 12)

Stolc V. and et. al. previously published a dataset that used microarray with splicing junction probes (SJPs) to detect splicing events over six developmental stages of Drosophila melanogaster[48] – embryo (0-2 hours), embryo (3-16 hours), larva, pupa, male, and female. The experiments consist of 24 dual-channel microarrays, which include 12 pair two-stage comparison (e.g. male v.s. female) and each pair has two experimental replicates. The data contains the intensity of hybridization of each channel for each SJP, and the relative intensity between the two channels approximates the relative

abundance of the corresponding 5'-3' junction happening in one sample versus the other. The microarray design also included a set of negative control probes (WJPs as named in the original paper[48]), which are constructed by swapping the SJP sequences in a way such that there should be no biologically meaningful hybridization. These WJPs were used to estimate the background signal level.

We chose this dataset because Drosophila is an excellent model system for studying alternative splicing, to which cutting-edge genetic and tissue culture technologies allows efficient experimental verification of computational results, and the majority of the genes involved in the regulation of alternative splicing were suggested to be shared between humans and flies, so that in principals what is learned by studying this process in Drosophila could be directly applicable to the regulation of alternative splicing in humans.

2.3.2 Statistical analysis of the SJP-microarray data

To obtain the information regarding the splicing profile of experimentally covered genes, which would be used as the input for the expert learning algorithm on the regulatory mechanisms, we processed the raw SJP-microarray data through the following major steps (Figure 13): (i) ANOVA analysis was performed to filter the differentially hybridized SJPs between the two samples within each pair, and the resulting p value was adjusted by the BY approach [19] to achieve a false discovery rate; (ii) controlling FDR $\leq$ 0.05, the significance level of differential splicing between two stages was projected onto one of the three cases {1, 0, -1} for representing {significant higher splicing

rate, no significant difference, significant lower splicing rate}, which results a differential splicing matrix (Figure 13); (iii) the expression level of the flanking exon (from the exonic probe microarrays of the same experimental setup) was combined to distinguish the missing transcription from missing splicing, which also provided the state of A.Site.E (the availability of the pre-mRNA) in the Bayesian network model of splicing regulation (see the Module Expert below); (iv) a voting method was taken for summarizing a single splice state for each splicing junction at each developmental stage based on the processed results over all the experiments that contains this developmental stage; (v) finally, the splice site level result from step (iv) was integrated with the gene structure annotation to achieve a gene level result, and only the alternatively spliced sites that are supported by both the annotation, which is mainly based on the EST data, and the SJP array are taken as the "trusted" alternatively spliced sites and are ported to the learning algorithm.

2.3.3 Overview of the Experts Learning (EL) algorithm for studying alternative slicing regulation

As the name infers, the EL algorithm consists of multiple experts, each of which is specialized in learning certain type of feature/regulatory relationship, which contributes to the regulation of alternative splicing. The goal of the algorithm is to learn the splicing regulatory modules (groups of splice sites that are sharing the same regulatory mechanism) and the multi-level cis-regulatory code (a {PWM + constrains} form representation of cis-regulatory information) at the same time. The goal is realized through iteratively calling the experts, leveraging the learning results among

experts for better (i.e. close to global optimal) choice of the initial state of a learning process, and tuning each experts' results for best simplest hypotheses of the regulatory mechanisms that fit the experimental observation. In the following, we firstly explain the probabilistic models underlie each expert, and describe the iterative learning algorithm that brings the expert's learning processes together.

The module expert uses probabilistic models to represent the relationship between biological domains, so that form a graphical representation in which the nodes represent relevant biological domains and the directed edges, with which a probabilistic model associated telling the probability of certain state of the children node based on the state of the parent node, represent the relationship between the linked nodes (a Bayesian Network). The Bayesian network was designed on top of Eran Segal's framework for transcriptional regulatory modules [59], as the mechanism for both transcriptional and splicing regulation involve the same kind of interaction between trans-factor and cis-codes. However, a regulatory module of alternative splicing has its special properties, which requires modification on the original network: firstly, the targeting units of the splicing regulation are the individual splice sites, instead of genes; secondly, the splicing regulation works on pre-mRNAs, so that the information regarding the availability of the pre-mRNA (i.e. the gene is transcribed) is need to be integrated into the model; thirdly, unlike the transcriptional regulatory motifs, which usually are located in promoter regions, the affective region (i.e. the region where a functioning regulatory motif is located) is different – here we made an assumption that the splicing regulatory motif is located

within the flanking exon or intron of a splice site.

Figure 14 illustrates the Bayesian network for the regulation of alternative splicing modules, in which we see the module expert portion (blue box) involves the following biological domains (the nodes): the motif profile of this affective region (Site.R), the splicing regulatory modules (i.e. groups of genes that are sharing the same regulatory mechanism) (Site.M), the temporal/spatial context (in this study, these are the six-developmental stages of Drosophila) represented by processed microarray result, which tells the availability of the pre-mRNAs (A.Site.E), and the splicing status at the splice sites (Site.A.C) in a cellular context. Associated with the directed edge "Site.R -> Site.M" and "A.Site.E -> Site.A.C" is the parameter u_{mr} and v_{am} respectively, both of which have a real-world meaning (see Equation 1 & 2). Given the state of Site.R, which would be calculated by the Motif Expert (see description under the Motif Expert), the state of Site.M is defined by Equation 1:

$$\Pr(Site.M \mid Site.R) = \prod_{site} \Pr(site.M = \bar{m} \mid site.R = \{r_1, r_2, ..., r_L\}) = \prod_{site} \frac{\exp\{\sum_{i=1}^{L} u_{\bar{m}i} r_i\}}{\sum_{m'=1}^{K} \exp\{\sum_{i=1}^{L} u_{m'i} r_i\}} \quad \text{(Eq. 1)}$$

where $r_i \sim$ a binary number representing whether site's flanking region contains cis-code *i*; there're *L* cis-code in total.

$u_{\bar{m}i} \sim$ the weight vector (length *K*) specifies the extent to which the cis-code *i* plays a regulatory role over all the module $\bar{m}$.

$u_{m'i} \sim$ the weight (a scalar) specifies the extent to which the cis-code *i* plays a regulatory role over the module *m'*. (*K* modules total)

Assuming that a cis-regulatory code is only contributing to the regulation of a small subset of the total modules, we limited the number of $u_{m'i}$ that are non-zero (h) to be << K, which results in a sparse weight matrix umr. In addition, we require all weights to be non-negative ($u_{m'i}$ >=0), intuitively which means that a gene's assignment to certain modules can only depend on the presence, but not absence, of a cis-regulatory code in the affective region. Note that these constrains on umr reduced the freedom of the parameter space, which helped to avoid overfitting during the learning step. Similar assumptions were also made in Segal's work on transcriptional module analysis.

Based on the resulting Site.M from Eq. 1 and the state of A.Site.E directly provided by the preprocessing of the exonic microarrays (see methods above), the state of Site.A.C could be calculated according to Equation 2:

$$\Pr(Site.A.C \mid Site.M, A.Site.E) = \prod_{site} \prod_{a} \Pr(site.a.C \mid site.M, a.site.E) \qquad \textbf{Eq. 2}$$

where $$\Pr(site.a.C \mid site.M, a.site.E) = \begin{cases} \sum_{i=1}^{K} site.m_i \cdot v_{ai} & ; \text{if } a.site.E = true; \\ 1 & ; \text{if } a.site.E = false; \end{cases}$$

$site.m_i$ ~ the relative strength of *site* being associated with module m_i; $\sum_{i=1}^{K} site.m_i = 1$;

v_{ai} ~ the normalized activity level of m_i's corresponding trans-factor in the array condition a; $v_{ai} \in [0,1]$

The intuition behind Equation 2 is that the splice state of site in sample condition a (site.a.C) only depends on what module/modules the site belongs to (site.M) and whether the gene was transcribed at the sample condition a (a.site.E). If the gene containing the site is not transcribed in the array a (i.e. a.site.E = false), there is no substance for splicing regulation, thus no contribution to the parameter estimation, for which we set $\Pr(site.a.C \mid site.M, a.site.E) = 1$; as long as the corresponding gene is transcribed in array a (i.e. a.site.E = true), the splice state of site in array a (site.a.C) is determined by the module assignment site.M and the activity level of each m_i in the array condition a (parameter v_{ai}). Note that to limit the parameter space, so that help avoiding overfitting, we constrained the parameter v_{ai} to be non-negative, which means the model was only constructed for the positive regulators. However, in most cases, a negative cis-regulatory program could still be identified by this model, because (i) the presence of a splicing repressor in some sample condition can always be learned as the absence of an enhancer, and similarly, the model can use the presence of a splicing enhancer in some

sample conditions to fake the absence of a splicing repressor in the same sample conditions. Thus it is dangerous to draw the conclusion about the positive/negative regulatory function of a learned cis-code directly from the parameter v_{am}; backgrounds on the biology of a particular case or experimental verification is needed for gaining better insights about this. Nevertheless, since a regulatory module tends to function in a particular set of conditions, if the learned parameter suggests that a module has a broad activity and only in a small set of conditions its activity is missing, this might suggest that the module has a repressive function over that small set of conditions.

In summary, the module expert predicts the splice states based on the parameter u_{mr} and v_{am}, the cis-code content, and the transcription state. The predicted splice state could be compared to the observed splice state from the SJP-microarray results to provide a "distance" to the experimental observation, which we take as the "correct answer" for the prediction, and such "distance" could direct "which way" and "how much" the parameters should be adjusted for achieving a better prediction result. This "adjusting" method is essentially a major step of the learning algorithm, which we described later in this section (See Learning Algorithm).

The Motif Expert completes the Bayesian network illustrated by Figure 14, by providing a model to calculate the state of Site.R based on the affective sequences of splice sites (Site.S) and a collection of cis-regulatory codes in the {PWM + constrain} type representation (the mapping function of the Motif Expert); it also plays a role in

learning, ab initio, PWM from a subset of Site.S that share the same cis-code r (the learning function of the Motif Expert).

Different from the common representation of a cis-regulatory code which limits the characteristic information within the motif sequence, our model represents a cis-code as a combination of both the motif sequence feature and the relevant statistical features beyond the motif sequence itself (e.g. spatial information, and more interestingly for splicing research, the information on the basal splicing signals), thus correspondingly, the mapping function of the Motif Expert also need to test for both criterions.

The motif sequence feature is represented with a positional weight matrix (PWM) and a cutoff score for significant matching scores. A PWM specifies the frequency distribution of nucleotides at each position of the binding sites and is considered to be related to the energy of binding of a trans-factor to the DNA/RNA[60]. The Motif Expert scans through a site.S, and for each window (with length equal to the length of the motif), it summarizes a matching score that tells the likelihood of this window being a significant match to the PWM represented motif sequence in consider the background frequency of each nucleotide (Eq. 3). A matching score which is greater than the cutoff value suggests a significant match to the mapping motif.

$$score_i = \sum_{j=1}^{P} \omega_j \left[S_{i+j-1} \right] \qquad \text{Eq. 3}$$

where $score_i \sim$ the matching score for the window starting at the i^{th} nt of a sequence;

$\omega_i[x]$ ~ the log-likelihood (the weight) of having nucleotide x at the i^{th} position $x \in \{A,T,G,C\}$

By integrating constrains into the cis-code representation, we gained a much richer model for cis-regulatory code. Constrain is specified as "measurements + N×[cutoff value + direction (i.e. > or <)]", where $N \in \{1,2\}$ which is the number of the "cutoff value + direction" pairs. Given a defined constrain, the mapping function of the Motif Expert calls the measurement, and compares the resulting value to the cutoff value to check if the direction is satisfied. For example, if the measurement is "distance to 5'ss", the cutoff value is "35 nts", and the direction is "<", and we have a candidate motif which passed the PWM scoring and is located 20 nts away from the 5'ss, then this candidate motif can pass the constrain test. Sometimes constrains can contain two "cutoff value + direction" pairs, which define both the upper and lower bound of a value.

Only when the region in the affective sequence passed both of the two-step test (PWM scoring test and the constrain test), will the Motif Expert accept it as a putative cis-code and update the corresponding site.R value.

In addition to the mapping function, the Motif Expert also performs a role in learning the sequence feature of a cis-code (i.e. only the PWM representation); it finds, ab initio, a PWM that is significantly enriched in a positive group of sequences in comparing to a null group (control group) using the random projection algorithm[61], which was previously published and has been broadly used. During overall iteration of learning,

when the learning function of the Motif Expert is called, it will take the current state of Site.R and put all the sites that are supposed to share the same regulatory mechanism into a positive group, and correspondingly pull together an equal number of randomly chosen 5' or 3' flanking sequences into the control group (i.e. the control group could contain the flanking exon + intron for both alternatively spliced or constitutively spliced sites), and call the random project-based motif finding algorithm.

The Constrain Expert handles the learning of cis-regulation-related information that is not captured by the PWM-only representation of a cis-code. According to the current biological model of the cis-regulation of alternative splicing, such information could potentially be the spatial information (e.g. the location of a motif in relative to the targeting basal splice signal), the combinatorial information (e.g. the synergy between motifs), or certain property associated with the targeting basal splice signals themselves. With this broad range of possibilities, adding constrains into the model for cis-regulatory code dramatically increased the complexity of the model, so that post a challenging computational question of how to learn statistically significant constrains from a limit amount of data without overfitting the model.

To limit the space for searching biologically relevant models, the Constrain Expert requires an explicit definition about the "measurement", which the considering constrains are based on. This measurement is the same one as appearing in the constrain representation described above. Given a defined measurement, the Constrain Expert

applies the Kolmogorov-Smirnov test (K-S test) to evaluate if the positive group and the control group come from the same distribution on this measurement (the null hypothesis H0) in considering the sample size. When the K-S test suggests rejection of the null hypothesis H0 (i.e. the positive group and the control group come from different distribution according to the particular measurement), the Constrain Expert calls a maximum likelihood estimator to learn the {cutoff + direction} part of the constrain representation. More specifically, from the empirical cumulative function of the two distributions, the Constrain Expert could figure out if the difference between the two distributions is on the mean, or the variance, or both, base on which the Constrain Expert chooses to learn only one cutoff (if the means are different) or two cutoffs (when the means are similar, while variances are different, two cutoffs were learned to specify a range of the value in the form [a, b]). Figure 15 gives the graphical illustration of these two different situations.

The design of the Constrain Expert allows the users to define multiple measurements, and because the K-S test has the advantage of making no assumption about the underlying distribution (i.e. different from the commonly used t-test that requires normal distribution, especially for small sample size) and being insensitive to the log-transformation, it provides a powerful uniformed platform for screening biologically meaningful models under different definitions. However, as any other statistical approaches, there's always a possibility of making a mistake, although the chance is small, associated with the conclusion; thus, a small set of carefully selected candidate measurements are

preferred than trying on a large number of randomly chosen measurements. According to the current biological model of alternative splicing regulation, we choose the following as the measurements for the Constrain Expert's consideration: (i) the intron/exon bias, which is represented as a binary valued constrain telling whether the cis-code is functioning as an exonic motif (in our constrain representation, the "cutoff" is used to store the binary value, and the "direction" is not used for this measurement); (ii) the log-transformed distance to the 3' or 5' splice site (the constrain representation as described in Motif Expert); (iii) the U1 affinity of 5'ss, which we measured as the state of the [-3, +6] positions (the 3 nts into the exonic side, and 6 nts into the intronic side of the 5'ss) being complementary to the corresponding nucleotide in U1 snRNA (this results a length 9 vector with binary entries for each site.S that a 5'ss is covered; note that this constrain is not applicable to the 3'ss).

The overall learning algorithm puts together the experts as a team specialized in learning alternative splicing related cis-regulatory code. Figure 16 summarized the iterative algorithm in diagram. The learning proceeds in two levels of iteration. The lower level iteration optimizes the parameter u_{mr}, v_{am} and the state of Site.R and Site.M in the Bayesian network to minimize the negative log-likelihood of that the learned network explains the observed data A.Site.E and Site.A.C; the top level iteration optimizes the cis-regulatory code, which are represented in the form of {PWM + constrains} by calling the learning function of the Motif Expert and the Constrain Expert; the learnt collection of cis-regulatory code is used to calculate a new state of Site.R by calling the mapping

function of the Motif Expert, and starts a new round of the lower level iteration; the iteration continues until little adjustment is made on the parameters and the state of Site.M between iterations.

The lower level learning is realized through the previously established algorithm for learning hidden regulatory modules for transcriptional regulation[12]. The algorithm fall into a general family of technique for statistical estimation, called Expectation and Maximization (the EM algorithm)[62]. In the Expectation step (E-step), the algorithm greedily finds the hidden state of Site.M and Site.R; in the Maximization step (M-step), the program makes forward prediction on the state of A.Site.C based on the current state of parameters and module assignments (Site.M), and according to "how far" the prediction is from the experimental observation for A.Site.C, it goes backward to adjust the parameters; the alternation between the E-step and M-step continues until the adjustment becomes very small between iterations. The details of the E-step and M-step were well explained in the Segal's paper[12] on learning modules of transcriptional regulation, and the differences between the previous work and our model (e.g. we introduced an extra node (A.Site.E) for representing the availability of the pre-mRNAs) were emphasized above in the description of the model design. It is noteworthy that because the modularity is a general property of biological regulation, and the trans-factor/cis-code-based regulatory mechanism is also shared between the alternative splicing regulation and the transcriptional regulation, the architecture of the Bayesian network and the constrains on the parameter space, thus also the learning

algorithm, is very similar to Eran Segal's design for studying transcriptional regulation.

However, in order to well address the complexity derived from the multi-level information associated with the functionality of the splicing cis-regulatory code, we introduced a novel platform for identifying the cis-code-related information that was not able to be captured by the PWM scoring model, and integrated its learning process with the learning of the modularity. This novel component is learned within the top level iteration (the details of the algorithm was described in the Motif Expert and Constrain Expert above), and the results are used for the re-initiation of the lower level iteration. Note that a cis-code can be added when it is found to be capturing the traceable sequence-level signature that can be used to predict the modularity and correspondingly the splicing regulation; meanwhile, a cis-code can also be eliminated during the learning, if without it there is no significant drop in the quality of prediction; moreover, a cis-code could be only partially modified, which not only includes a change in the PWM, but also can be the change/deletion of the constrains within the cis-code representation.

2.4 Results and Discussion

We chose to experiment our approach on the previously published SJP-microarray dataset[48] on six developmental stages of Drosophila melanogaster (see Methods and Materials for details). These SJP-microarrays detected 3,955 predicted genes with as few as 2 and as many as 54 exons[48], among which we identified 2542 (64.3%) of the genes having a hybridization signal level above the background by comparing to the signal

level of the WJP probes, in at least one of the microarrays. This number is very close to the previous analyzing result (2606 (65.9%) genes). Through the five-step pre-processing of the SJP hybridization signal (the raw data from) we identified 75 genes showing significant alternative splice forms across the six different developmental stages, which is a strikingly small number comparing to the previous estimation that 53% percent of the expressed genes (1374 out of 2606 genes) exhibited exon skipping.

We carefully compared our pre-processing methods (see Methods and Materials) to the previous analyzing methods[48], and realized there are at least two major difference in methodology that can explain the difference in results: First, we identified the alternatively spliced events by significant differential expression of SJPs between stages, which involves the ANOVA analysis, FDR control (we controlled FDR<0.05), distinguishing missing transcription from missing splicing, and a voting step that considers the multiple two-channel microarrays' result, which covered the same sample stage, to derive a single splice state for that particular developmental stage relative to all the other five stages (step i to iv in our pre-processing method). Comparatively, previous analysis on the SJP-microarray data was carried out at per-channel level and the signal was compared with the background noise that was measured as the signal level of the WJPs; thus, a SJP would be considered as alternatively spliced if in one/some, but not in all the sample stages, this SJP has an above background signal level. This method does not address if there's differential hybridization between the stages where the SJP shows above background signal level; it does not correct the p values for multiple hypothesis

test, which could dramatically increase the false positives; and it doest not distinguish the case of no transcription from the case of missing splicing. Second, to reduce the false positives, we required that the SJP-microarray suggested alternative splicing events have to be also supported by the annotation data, in which the alternative transcripts were annotated mainly based on EST data. We noticed that among the 2542 genes that have above background level hybridization in the SJP-microarrays, there are only 417 genes annotated as alternatively spliced genes, among which 18.0% (75 out of 417) showed significant differential splicing over the six developmental stages that the experiments covered. Comparatively, the previous analysis did not require support from an independent recourse. Actually, over one half of the alternative splicing events suggested by previous SJP-microarray data analysis did not have any EST support[48]. In summary, our method of preprocessing the SJP-microarray data minimized the false positive rate by guaranteeing an upper bound of FDR as 0.05 and an agreement between the array data and the annotation data, which potentially could cause a higher false negative rates than the previously result; however, the previous result is likely to contain large false positive rate. As a low positive rate was preferred for the purpose of studying the regulatory mechanism, we chose the alternative splicing profiles suggested by our approach as the input to the learning algorithm.

We applied the Expert Learning algorithm to the set of 185 alternatively spliced sites (alt-ss) from the 75 differentially spliced genes. As the EM algorithm that we used for learning the regulatory modules only guarantees finding a local optimal, if

multiple local optima exist, theoretically the learning results are sensitive to the initial state of the parameters and the arbitrarily chosen value of K, the number of modules to be learned. Since our dataset covers six developmental stages, a reasonable choice of K is 6, assuming each stage has its own stage-specific splice regulatory mechanism. For evaluating the robustness of our approach, we experimented with the value of K being 5, 6, or 7. We observed very similar negative log-likelihood (-LL value) at convergence for K=6 and K=7 (-LL is equal to -1.68 and -1.82, respectively), and a big drop of negative log-likelihood (i.e. better fitting) when change K from 5 to 6 (-LL droped from -0.43 to -1.68). This suggested that 6 is a good choice for K, which made intuitive sense as our experimental data covered six developmental stages.

We plot the learned parameter v_{am} for the case K=7 to evaluate the potential overfitting in comparing to the case K=6 (Figure 17 C). We noticed that among the seven identified splicing modules, there are five showing significant specificity to certain developmental stage and For example, the module m_6 is likely to be female specific, module m_5 is male specific, module m_4 is specific to pupa stage, module m_2 is specific to embryo 3~16hour stage, and m_7 shows similar activity level for lava and male stage. The other two modules seem to function in most of the developmental stages, except for a difference in the early embryo stage (0~2hours); m_3 shows slightly higher activity level for the early embryo stage (0~2hours), while m_1 did not show any function for this stage. The broad activity of m_1 and m_3 shown in the parameter v_{am} at convergence suggested the possibility of

them being ubiquitously functioning mechanism for assisting splicing event at sub-optimal splice sites; in another word, these mechanisms do not show clear context-specificity, at least not for different developmental stages. Another possibility is that the modules have negative regulatory effects on the splicing. Because our model for splicing regulation only assumed positive regulatory effects in order to reduce the parameter searching space, the model lost the direct inference of negative/positive regulatory function from the learning results. However, as discussed in the Methods for the Module Expert, in our learning program, the negative regulatory mechanism would be identified as a positive regulator with activities in the complementary set of stages of the actual negative regulator (see Methods for details). Thus it's possible that m_1 and m_3 function as tissue-specific negative regulators (e.g. m_1 could have early embryo (0-2hour embryo) stage-specific negative regulatory function).

To distinguish these two possibilities, we compared the two cis-regulatory codes (r_1 and r_3) (Figure 17A), which were learned by our program as exon specific motif and were suggested to function in m_1 and m_3, to the known splicing motifs stored in ESE Finder[63]. We found that r_1 contains human SF2/ASF binding sites and r_3 contains human SRp40 binding sites. Both of these regulators were previously known as commonly expressed essential splicing factors that can enhance splicing through interaction with exon specific pre-mRNA motifs, and recently experimental evidences indicated that they affect alternative splicing through a concentration-dependent

manner[64,65]. Thus, it is likely that the identified r_1 and r_3 also function as positive splicing regulators over broad developmental stages in Drosophila.

However, this potential common functionality does not suggest that they do not have tissue-specific function. For example, module m_3, to which both r_1 and r_3 have significant contribution, demonstrated changes in activity level over the six developmental stages. More specifically, it has the highest activity level in the early embryo development (0-2hour) stage, while only about half activity level over the later stages. This is especially interesting, since the SF2/ASF was known to be essential for embryonic development, upregulation of SF2/ASF was identified in various human tumors, and its over-expression is sufficient for inducing transformation in cell-line studies[65]. It is possible that this function in conserved from fly to human. Moreover, the exonic distribution of r_1 and r_3 also supported their potential role as both essential splicing factors and splicing regulators through changes in concentration; both of these two motifs appears in many alternative splice sites' adjacent exons in compare to more tissue-specific cis-code, but a large variety in copy numbers was also observed(Figure 17B), and combinatorial effects between these two cis-code was not only suggested by the shared contribution of r_1 and r_3 to m_3 but also supported by their co-appearance in multiple exons. Thus, multiple evidences strongly suggested that r_1 and r_3 are biologically meaningful cis-regulatory code for alternative splicing and their contributing splice modules, although demonstrated broad activity over different developmental stages, match the known

functional properties of their potential trans-factors (SF2/ASF and SRp40), thus are likely pointing to meaningful splicing regulatory mechanisms.

We next examined the cis-regulatory code that were learned to be contributing to the regulation of developmental stage-specific splicing modules (Figure 18). There were two exonic and two intronic motifs were identified with such stage-specific property, two of which demonstrated significant similarity to previously known splicing regulatory motifs, and for two of these three, our program suggested a stage-specific regulatory role that was supported by previous biological evidences. More specifically, a T-rich intronic motif was identified to be female-specific cis-regulatory motif and it highly resembles the sequence feature of the dsx binding site, an intronic motif that were known to regulate female specific alternative splicing on genes related to fly sex determination. Moreover, an identified exonic motif with significantly high contribution to 3-16hr embryo development contains the binding signature of a human SR protein, SRp40 (r_2 in Figure 17). Previous experimental studies suggested SRp40's role in regulating alternative splicing in multiple biological contexts in human. For example, SRp40 was indicated to regulate insulin triggered alternative splicing by changing its phosphorelation states[66], and in human myometrial cells, SRp40 demonstrated regulatory role in changing the splicing form of a developmentally critical transcriptional factor, CREM, so that modifies the transcriptional factor from a transcriptional activator to a transcriptional repressor[67]. There has been no clear evidence of a similar SR protein in *Drosophila*. However, as many splicing regulatory proteins are conserved from fly to human, the

computationally identified SRp40 like cis-element is likely to be a biologically meaningful motif, and the computationally suggested regulatory specificity to late embryo developmental stage provided intriguing direction for searching the related transfactors.

Furthermore, a G-rich cis-element was identified with specific contribution to the pupa stage (r_4 in Figure 17). Although there has been known splicing regulatory motifs showing strong similarity, this G-rich element demonstrated very interesting location bias and 5' U1 affinity feature. More specifically, the Constrain Expert of the program identified a positional constrain for this G-rich motif as being located within 38 nts upstream of the corresponding 3'ss. This location bias is fairly intriguing, because this pro-3'ss region is very often U/T rich, which could cause very different local structural property than a G-rich pro-3'ss sequence, as the former provide flexible RNA chain, which is could potentially benefit the second catalytic reaction during splicing, while a G-rich element could dramatically increase the stiffness of the local RNA sequence so that affect the splicing reaction. According to this hypothesis, this G-rich element could by itself be a repressive element to the splicing of the adjacent 3'ss; however this does not prevent the possibility that a positive transfacor can bind to this element, so that suppress the negative effect and enhance the splicing at the downstream 3'ss.

Most importantly, the G-rich element consiste of a second constrain, which refers to the 5'ss U1 affinity. This is a rather surprising result as the 5'ss U1 binding strength was

often considered as a criterion for the 5'ss splicing, and the splicing state on a weak 5'ss was often considered as under exonic motifs, most commonly SR protein binding sites' regulation. However, when we took a closer look at the learned 5'ss property, we found that it refers to the distribution of the U1 finding nucleotide, instead of the overall U1 affinity. Comparing the genes within the pupa stage-specific module, which are with or without this G-rich element, we identified highly significant difference in the distribution of the U1 binding nucleotide (Figure 19). The G-rich element-containing genes' corresponding 5'ss demonstrated significantly higher than normal binding affinity at the intronic position +3 and +4 and significantly weaker than normal binding affinity at exonic positions -3~-1 (Figure 19B), while the other pupa stage splicing module members showed the opposite U1 binding characteristic (Figure 20A). It is noteworthy that both case demonstrated similar overall U1 affinity level, which suggested that the differential effects on the splicing decision might not be the recruitment of U1 associated spliceosome onto the 5'ss. Instead, combine the hypothesis of the G-rich element's role in affecting the second step of the splicing reaction, as discussed above, we favor a hypothesis that the 5'ss U1 binding property is related to the selection of the downstream 3'ss. In this sense, the basal splicing signals (i.e. the 5'ss splice sites) might also play a regulatory role. Actually, emerging experimental studies have suggested the selective role of certain 5'ss on the splicing of the downstream 3'ss. However, there has been no clear hypothesis that can explain why some of the similarly suboptimal 5'ss can carry different selective information regarding the downstream 3'ss splicing. Our computational results pointed to a role of the intronic +3/+4 and exonic -3~-1 positions' relative U1 binding

property in affecting the second enzymetic reaction of splicing events, and proposed promising direction for further experimental validation.

In conclusion, we presented a novel approach, which consists of a mixture of "experts" that are specialized in learning a partition of alternatively spliced sites into regulatory modules, learning the module-specific short sequence signatures (motifs), and learning the biologically meaningful constrains associated with each motif (such as motif combination, motif spatial distribution and basal splice signals), and developed an algorithm can effectively learn the model parameters from publicly available splicing detection data (e.g. splicing-related microarray data), which iterates over the experts and optimizes each expert's learning results for the simplest best overall explanation of the experimentally observed alternative splicing profiles. Using Drosophila developmental dataset generated with splicing-junction probes[48], our approach identified both previously known and unknown cis-regulatory motifs, associated their regulatory function with specific splicing profiles, and more importantly, for some motifs, our program selected intriguing higher level properties that could be critical in controlling the specificity of the motifs' regulatory function.

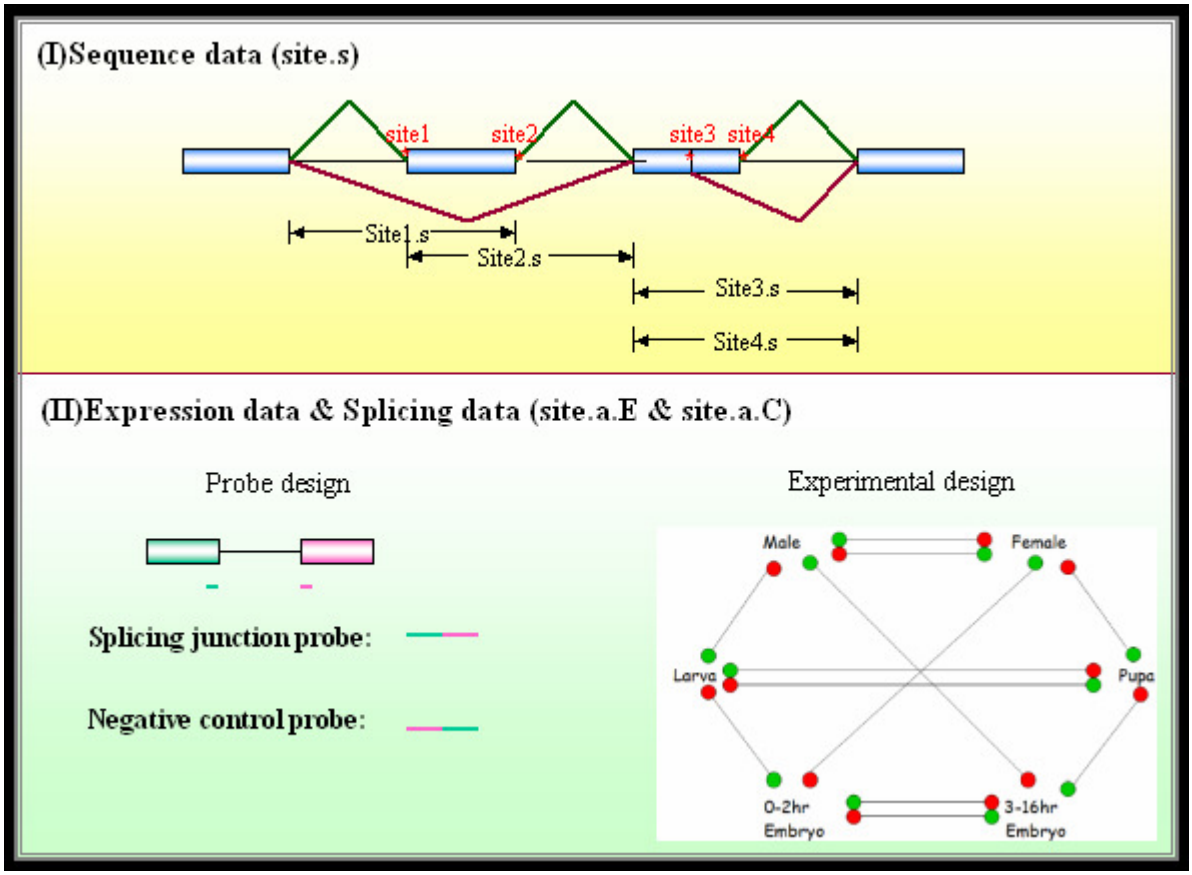


Figure 12. The input data for Mixed Expert learning. (I) illustrate the sequence data from the annotation resource and the cDNA based annotation of alternatively spliced sites (the red asterisk labeled splice sites). The lower penal shows the regions chosen to be the affective regions that are flanking an alternatively spliced sites. (II) shows the experimental design of the splicing-junction probes (the left) and samples for the two channel microarray experiments (the right)[48].

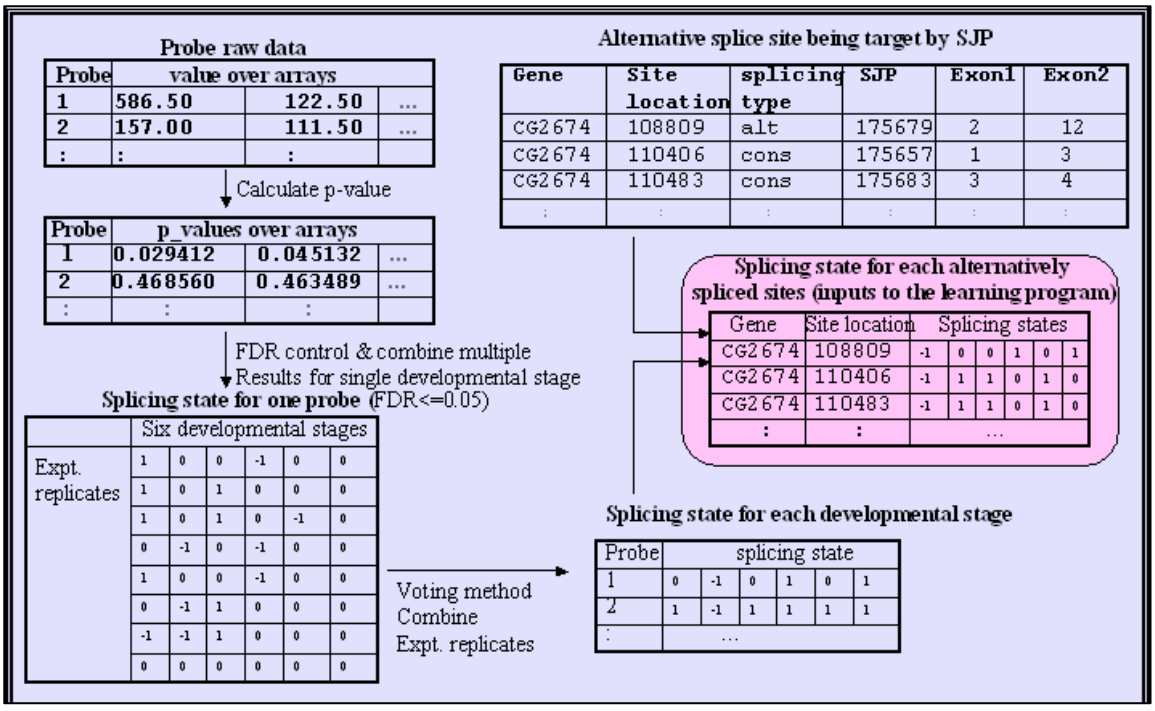


Figure 13. Detecting alternative splicing events using the SJP-microarray data. The diagram illustrates the five-step pipeline for pre-processing the SJP-microarray data to achieving a splice map of each probed splice site over the six developmental stages. The results were used as the input to the Expert Learning algorithm.

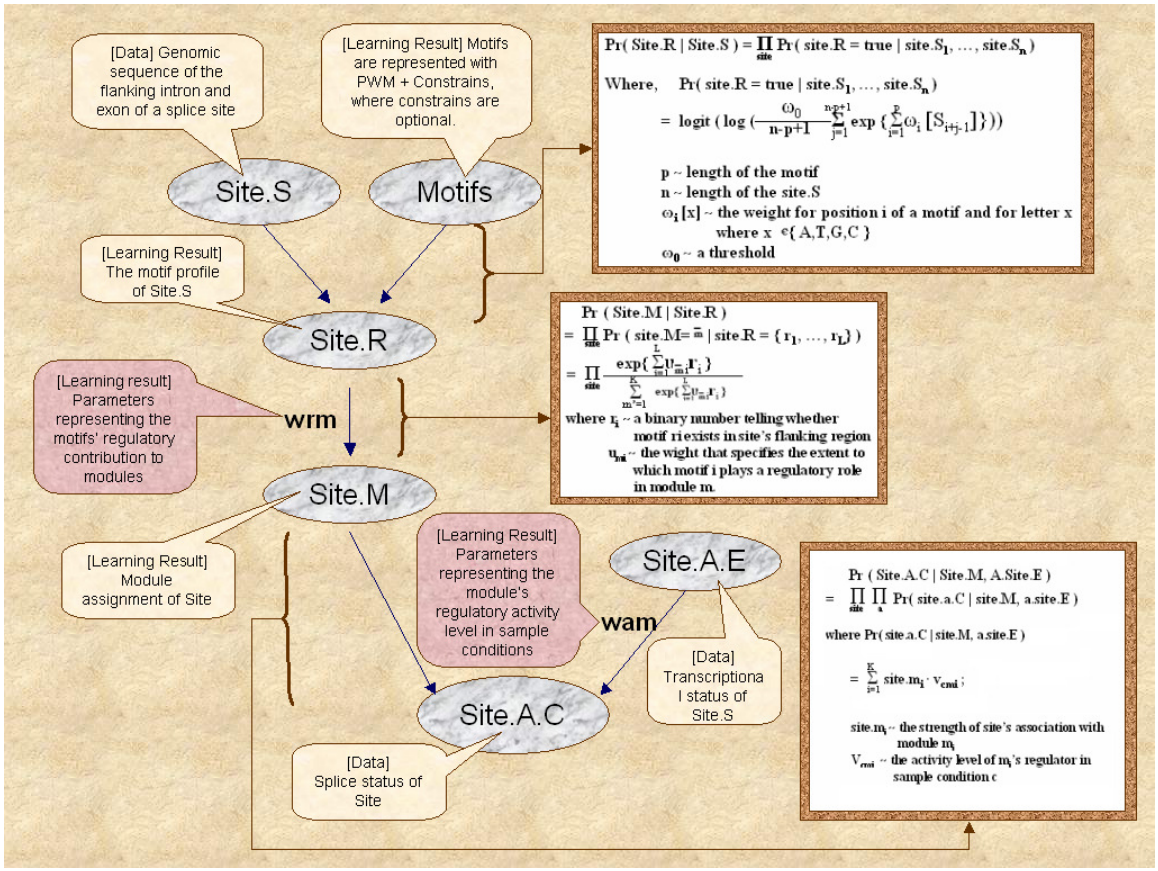


Figure 14. The Bayesian network and relational probabilistic models splicing regulation (Details described in Methods).

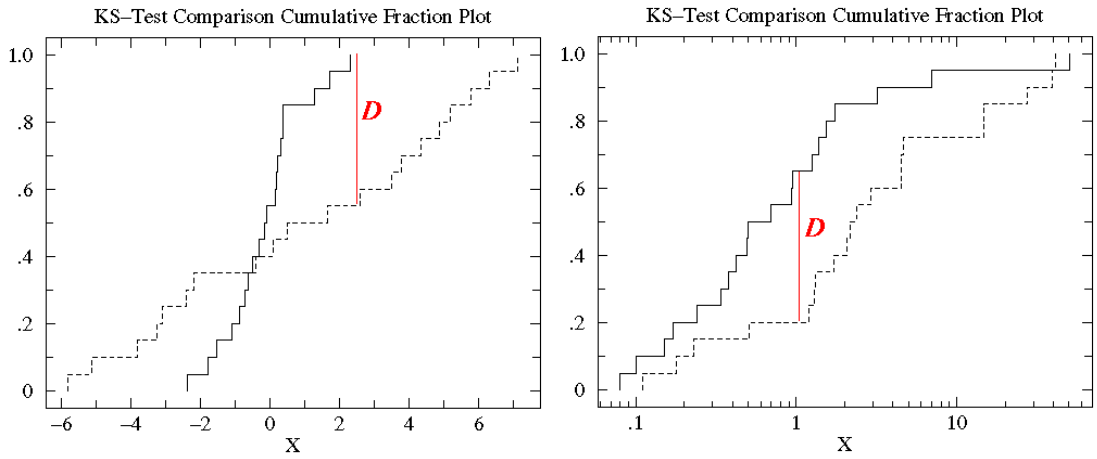


Figure 15. Graphical representation of KS-Test comparison cumulative fraction. The two plots show the empirical cumulative functions of certain measurements for the testing sample group (solid line) and the control group (dot line). (A) shows the case when the sample group demonstrated a significantly smaller variance while a similar mean to the control group, for which two cutoff scores are learnt (the lower bound and upper bound) for describing the sample group specific distribution. (B) shows the case when the sample group is more highly populated at the lower values compared with the control group. In this case, only one cutoff (the upper bound) is learnt. Similarly a single lower bound would be learnt if the sample group is more highly populated at the higher values.

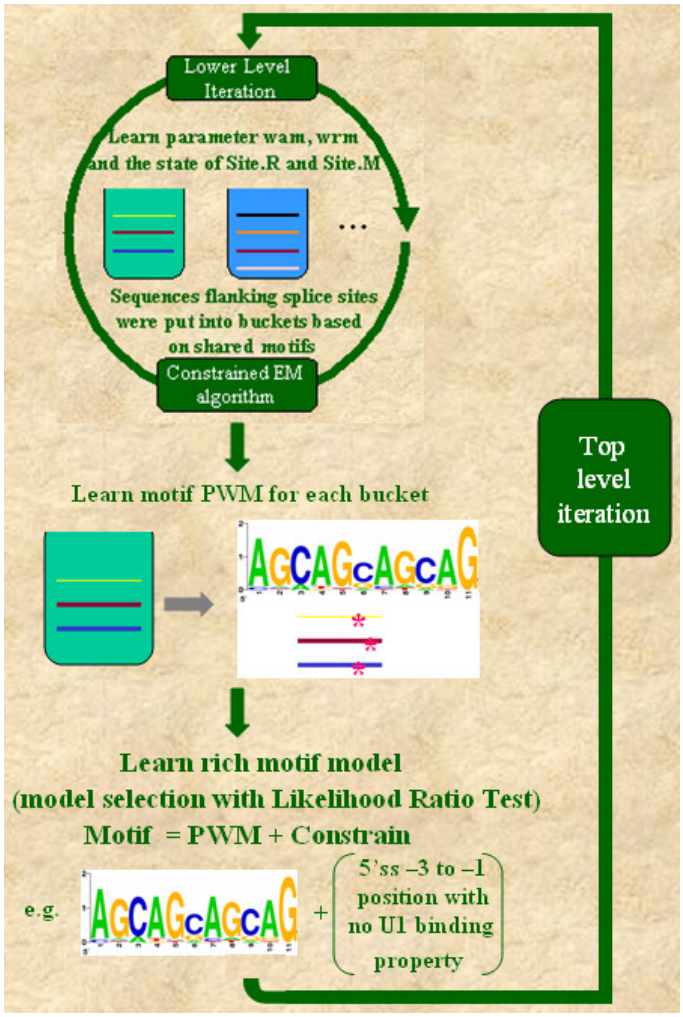


Figure 16. The iterative algorithm of Expert Learning.

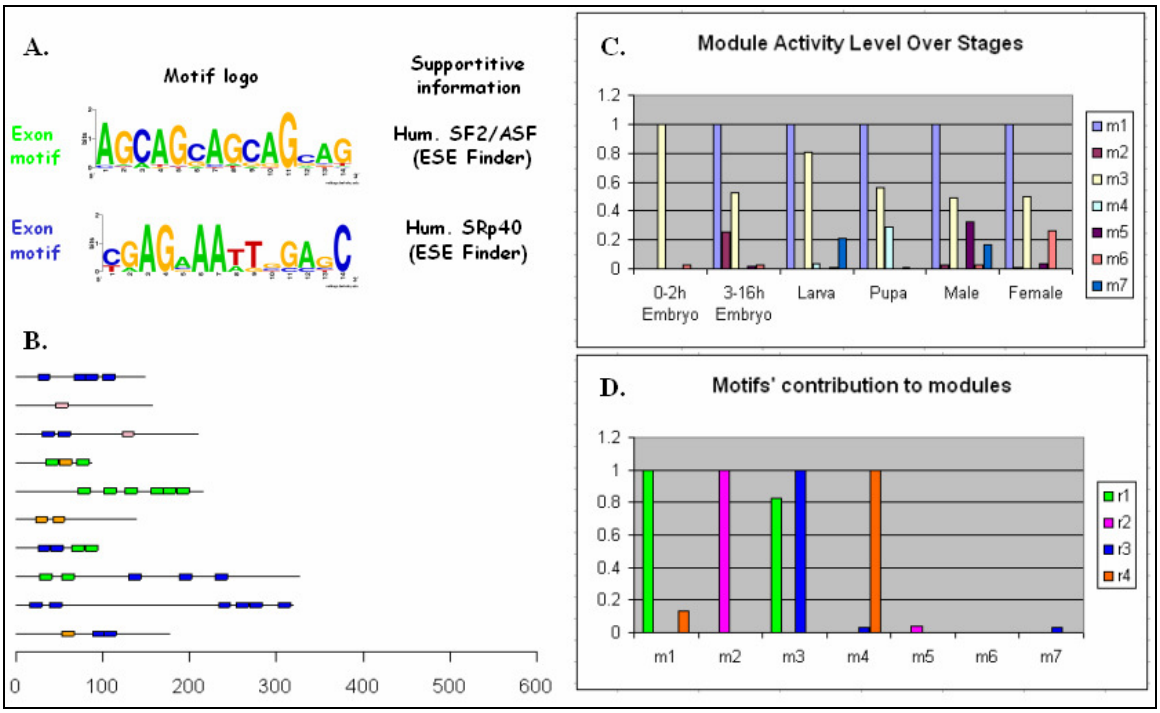


Figure 17. Two identified exonic motifs that could not show clear developmental stage specificity. (A) shows the Motif Logo of the two exonic motifs that are highly similar to the binding sites of two broadly expressed human SR proteins according to ESE finder[63]. The motifs are labeled with green and blue color, and the colors are consistently used in (B) and (D). (B) gives some examples of how the motifs distributed over the hosting exonic regions. Notably the blue and green motif shows higher copy numbers compared with the orange and pink motifs. (C) visualizes the parameter set V_{am}, which represents the regulatory activity level of each splice module over the six developmental stages for the case of module number $K = 7$. Five of the seven identified modules (namely m_2, m_4, m_5, m_6, m_7) demonstrate strong developmental stage specificity, while the other two (m_1 and m_3) shows broad functionality. (D) visualizes the parameter set U_{mr}, which shows how much each identified motif contribute to the seven learned splice modules. This histogram only covers four of the identified motifs for readability. The color label of the four motifs is the same as the one used in (B) and (A). Interestingly, the green (r_1) and blue (r_3) motif demonstrated high contribution to module m_1 and m_3, which are the two modules could not show stage-specific regulatory function, suggesting these two motifs are commonly functions in assisting weak splice sites' splicing and might not have developmental stage-specific role. This computational conclusion matches what is known about SF2/ASF and SRp40, whose binding sites show high similarity to the identified green and blue motifs (Details see text).

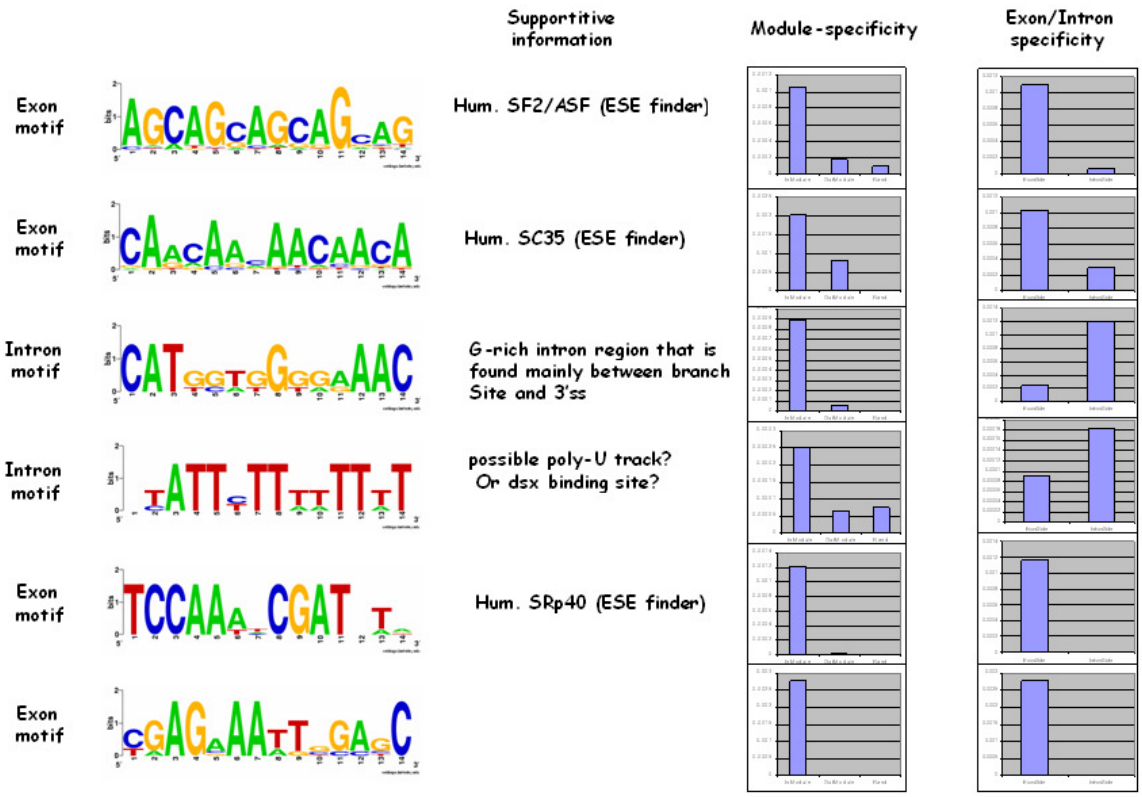


Figure 18. Identified splicing cis-regulatory code demonstrated significant module-specificity and exon / intron specificity. The left penal shows the Logo plots of the sequence feature of the identified motifs and their exon / intron specificity constrain, which was also learned by the program. The middle penal provides the supporting information found from literature. The right penal use histograms to summarize the module specificity and the exon / intron specificity; in Module-Specificity bar plots, the three categories in the histograms are the frequency of identifying a copy of the corresponding motif in (the left column) the module that the program suggested the motif contributes to, (the middle column) the sequences elsewhere in the genome, and (the right column) a group of randomly reshuffled sequences. In the Exon/Intron specificity bar plots, the left column shows the frequency of finding a copy of the corresponding in the exonic regions and the right columns shows the frequency in the intronic regions. All the frequency was measured in per nt.

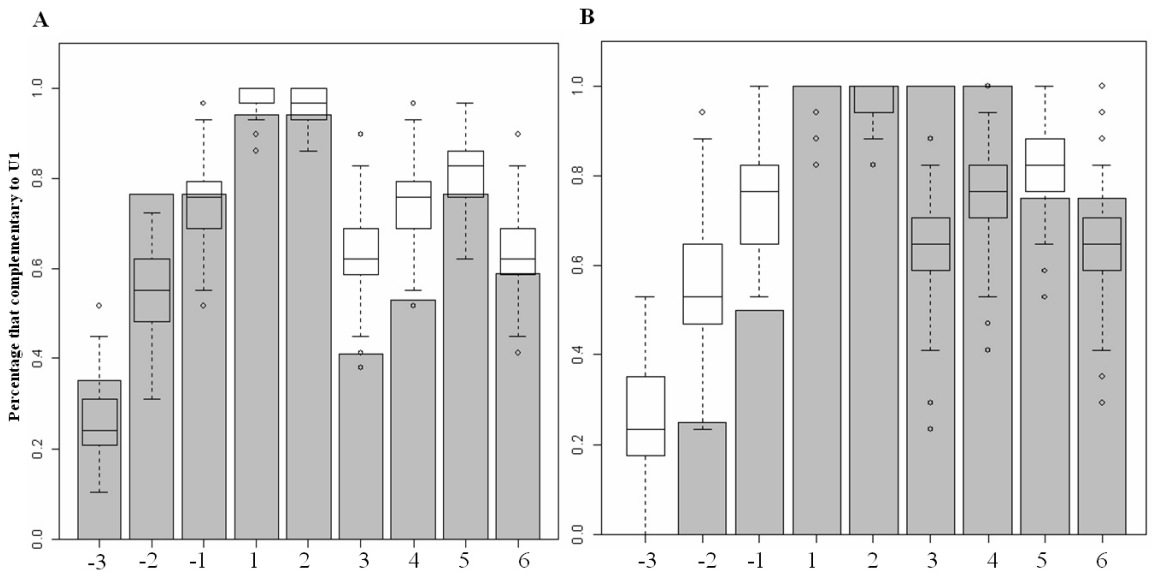


Figure 19. Identified module-specific 5'ss U1 binding property. Y-axis plots the frequency at which each nucleotide is complementary to the corresponding U1 nucleotide. The left penal and right penal shows the results over 5'ss within the pupa stage-specific splicing module that contain or not contain the intronic G-rich element respectively. The box plot shows the statistics estimated from randomly sampled sets from the total inputs, in which the random sample size is the as the size the module under examination.

2.5 References

1. Schmeichel KL, Bissell MJ (2003) Modeling tissue-specific signaling and organ function in three dimensions. J Cell Sci 116: 2377-2388.
2. Petersen OW, Ronnov-Jessen L, Howlett AR, Bissell MJ (1992) Interaction with basement membrane serves to rapidly distinguish growth and differentiation pattern of normal and malignant human breast epithelial cells. Proc Natl Acad Sci U S A 89: 9064-9068.
3. Koster MI, Kim S, Mills AA, DeMayo FJ, Roop DR (2004) p63 is the molecular switch for initiation of an epithelial stratification program. Genes Dev 18: 126-131.
4. Debnath J, Mills KR, Collins NL, Reginato MJ, Muthuswamy SK, et al. (2002) The role of apoptosis in creating and maintaining luminal space within normal and oncogene-expressing mammary acini. Cell 111: 29-40.
5. Jayanta Debnath JSB (2005) Modelling glandular epithelial cancers in three-dimentional cultures. Nature Reviews 5: 675-688.
6. Mailleux AA, Overholtzer M, Schmelzle T, Bouillet P, Strasser A, et al. (2007) BIM Regulates Apoptosis during Mammary Ductal Morphogenesis, and Its Absence Reveals Alternative Cell Death Mechanisms. Dev Cell 12: 221-234.
7. Connolly JL, Boyages J, Schnitt SJ, Recht A, Silen W, et al. (1989) In situ carcinoma of the breast. Annu Rev Med 40: 173-180.
8. Tsikitis VL, Chung MA (2006) Biology of ductal carcinoma in situ classification based on biologic potential. Am J Clin Oncol 29: 305-310.
9. Fournier MV, Martin KJ, Kenny PA, Xhaja K, Bosch I, et al. (2006) Gene expression signature in organized and growth-arrested mammary acini predicts good outcome in breast cancer. Cancer Res 66: 7095-7102.
10. Jayanta Debnath KRM, et. al. (2002) The role of apoptosis in creating and maintaining luminal space within normal and oncogene-expressing mammary acini. Cell 111: 29-40.
11. Debnath J, Muthuswamy SK, Brugge JS (2003) Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. Methods 30: 256-268.
12. Eran Segal YB, Itamar Simon, Nir Friedman, Daphne Koller (2001) From promoter sequence to expression: a probabilistic framework. technical report, Computer Science Department, Standford University.
13. Eran Segal NF, Daphne Koller, Aviv Regev (2004) A module map showing conditional activity of expression modules in cancer. Nature Genetics 36: 1090-1098.
14. Vamsi K Mootha CML, et. al. (2003) PGC-1alpha-responsive genes involved in xoidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics 34: 267-273.
15. Lucy Erin O'Brien MMPZ, Keith E. Mostov (2002) Building epithelial architecture: insights from three- dimensional culture models. Nature

Reviews Molecular Cell Biology 3: 531-537.
16. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
17. Virginia Goss Tusher RT, Gilbert Chu (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America 98: 5116-5121.
18. Gordon K. Smyth JM, Hamish S. Scott (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics 21: 2067-2075.
19. Yoav Benjamini DY (2001) the control of the false discovery rate in multiple testing under dependency. The annals of statistics 29: 1165-1188.
20. Michael Ashburner CAB, et. al. (2000) Gene Ontology: tool for unification of biology. Nature Genetics 25: 25-29.
21. Minoru Kanehisa SG, et. al. (2002) The KEGG databases at GenomeNet. Nucleic Acids Research 30: 42-46.
22. Kam D. Dahlquist NS, et. al. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nature Genetics 31: 19-20.
23. Anat Reiner DY, Yoav Benjamini (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19: 368-375.
24. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102: 15545-15550.
25. David A. Glesne ea (2006) Subtractive transcriptomics: establishing polarity drives in vitro human endothelial morphogenesis. Cancer Research 66: 4030-4040.
26. Caldwell GM, Jones C, al. e (2004) The Wnt antagonist sFRP1 in colorectal tumorigenesis. Cancer Research 64: 883-888.
27. Brisken C, Heineman A, Chavarria T, al. e (2000) Essential function of Wnt-4 in mammary gland development downstream of progesterone signaling. Genes & Development 14: 650-654.
28. Gattelli A, Cirio MC, Quaglino A, Schere-Levy C, Martinez N, et al. (2004) Progression of pregnancy-dependent mouse mammary tumors after long dormancy periods. Involvement of Wnt pathway activation. Cancer Res 64: 5193-5199.
29. Inadera H, Dong HY, Matsushima K (2002) WISP-2 is a secreted protein and can be a marker of estrogen exposure in MCF-7 cells. Biochem Biophys Res Commun 294: 602-608.
30. van Genderen C, Okamura RM, Farinas I, Quo RG, Parslow TG, et al. (1994) Development of several organs that require inductive epithelial-mesenchymal interactions is impaired in LEF-1-deficient mice. Genes Dev 8: 2691-2703.
31. Andl T, Reddy ST, Gaddapara T, Millar SE (2002) WNT signals are required for the initiation of hair follicle development. Dev Cell 2: 643-653.
32. Chu EY, Hens J, Andl T, Kairo A, Yamaguchi TP, et al. (2004) Canonical WNT signaling promotes mammary placode development and is essential for initiation of mammary gland morphogenesis. Development 131: 4819-

4829.
33. Simons M, Gloy J, Ganner A, Bullerkotte A, Bashkurov M, et al. (2005) Inversin, the gene product mutated in nephronophthisis type II, functions as a molecular switch between Wnt signaling pathways. Nat Genet 37: 537-543.
34. Brown AM (2001) Wnt signaling in breast cancer: have we come full circle? Breast Cancer Res 3: 351-355.
35. Polakis P (2000) Wnt signaling and cancer. Genes & Development 14: 1837-1851.
36. Weeraratna AT, Jiang Y, al. e (2002) Wnt5a signaling directly affects cell motility and invasion of metastatic melanoma. Cancer Cell 1: 279-288.
37. Weaver VM, Lelievre S, Lakins JN, Chrenek MA, Jones JC, et al. (2002) beta4 integrin-dependent formation of polarized three-dimensional architecture confers resistance to apoptosis in normal and malignant mammary epithelium. Cancer Cell 2: 205-216.
38. Trompouki E, Hatzivassiliou E, Tsichritzis T, Farmer H, Ashworth A, et al. (2003) CYLD is a deubiquitinating enzyme that negatively regulates NF-kB activation by TNFR family members. Nature 424: 793-796.
39. Brummelkamp TR, Nijman SMB, Dirac AMG, Bernards R (2003) Loss of the cylindromatosis tumour suppressor inhibits apoptosis by activating NF-kB. Nature 424: 797-801.
40. Mills AA, Zheng B, Wang XJ, Vogel H, Roop DR, et al. (1999) p63 is a p53 homologue required for limb and epidermal morphogenesis. Nature 398: 708-713.
41. Eddy SF, Guo S, Demicco EG, Romieu-Mourez R, Landesman-Bollag E, et al. (2005) Inducible IkappaB kinase/IkappaB kinase epsilon expression is induced by CK2 and promotes aberrant nuclear factor-kappaB activation in breast cancer cells. Cancer Res 65: 11375-11383.
42. James A. Mobley RWB (2004) Estrogen receptor-mediated regulation of oxidative stress and DNA damage in breast cancer. Carcinogenesis 25: 3-9.
43. Cadigan KM, Nusse R (1997) Wnt signaling: a common theme in animal development. Genes & Development 11: 3286-3305.
44. Iris Alroy YY (1997) The ErbB signaling network in embryogenesis and oncogenesis: signal diversification through combinatorial ligand-receptor interactions. FEBS Letters 410: 83-86.
45. Jie Feng AVR, et. al. (2004) Processing enzyme glucosidase II: proposed catalytic residues and developmental regulation during the ontogeny of the mouse mammary gland. Glycobiology 14: 909-921.
46. Schmelzle T, Mailleux AA, Overholtzer M, Carroll JS, Solimini NL, et al. (2007) Functional role and oncogene-regulated expression of the BH3-only factor Bmf in mammary epithelial anoikis and morphogenesis. Proc Natl Acad Sci U S A 104: 3787-3792.
47. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. PNAS 98: 5116-5121.
48. Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, et al. (2004) A gene expression map for the euchromatic genome of Drosophila melanogaster. Science 306: 655-660.
49. Novoyatleva T, Tang Y, Rafalska I, Stamm S (2006) Pre-mRNA missplicing

as a cause of human disease. Prog Mol Subcell Biol 44: 27-46.
50. Venables JP (2006) Unbalanced alternative splicing and its significance in cancer. Bioessays 28: 378-386.
51. Huang R, Huang J, Cathcart H, Smith S, Poduslo SE (2007) Genetic variants in brain-derived neurotrophic factor associated with Alzheimer's disease. J Med Genet 44: e66.
52. Glatz DC, Rujescu D, Tang Y, Berendt FJ, Hartmann AM, et al. (2006) The alternative splicing of tau exon 10 and its regulatory proteins CLK2 and TRA2-BETA1 changes in sporadic Alzheimer's disease. J Neurochem 96: 635-644.
53. Kamma H, Portman DS, Dreyfuss G (1995) Cell type-specific expression of hnRNP proteins. Exp Cell Res 221: 187-196.
54. Park JW, Parisky K, Celotto AM, Reenan RA, Graveley BR (2004) Identification of alternative splicing regulators by RNA interference in Drosophila. Proc Natl Acad Sci U S A 101: 15974-15979.
55. Caceres JF, Stamm S, Helfman DM, Krainer AR (1994) Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. Science 265: 1706-1709.
56. Matter N, Herrlich P, Konig H (2002) Signal-dependent regulation of splicing via phosphorylation of Sam68. Nature 420: 691-695.
57. Mabon SA, Misteli T (2005) Differential recruitment of pre-mRNA splicing factors to alternatively spliced transcripts in vivo. PLoS Biol 3: e374.
58. Kuo HC, Nasim FH, Grabowski PJ (1991) Control of alternative splicing by the differential binding of U1 small nuclear ribonucleoprotein particle. Science 251: 1045-1050.
59. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34: 166-176.
60. Stormo GD, Fields DS (1998) Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci 23: 109-113.
61. Buhler J, Tompa M (2002) Finding motifs using random projections. J Comput Biol 9: 225-242.
62. Heckerman D (1995) A tutorial on learning with Bayesian networks. . Learning in Graphical Models: 274-284.
63. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR (2003) ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res 31: 3568-3571.
64. Huang Y, Steitz JA (2005) SRprises along a messenger's journey. Mol Cell 17: 613-615.
65. Karni R, de Stanchina E, Lowe SW, Sinha R, Mu D, et al. (2007) The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat Struct Mol Biol 14: 185-193.
66. Patel NA, Kaneko S, Apostolatos HS, Bae SS, Watson JE, et al. (2005) Molecular and genetic studies imply Akt-mediated signaling promotes protein kinase CbetaII alternative splicing via phosphorylation of serine/arginine-rich splicing factor SRp40. J Biol Chem 280: 14302-14309.
67. Tyson-Capper AJ, Bailey J, Krainer AR, Robson SC, Europe-Finner GN (2005) The switch in alternative splicing of cyclic AMP-response element

modulator protein CREM{tau}2{alpha} (activator) to CREM{alpha} (repressor) in human myometrial cells is mediated by SRp40. J Biol Chem 280: 34521-34529.
68. Fuchs E, Raghavan S (2002) Getting under the skin of epidermal morphogenesis. Nat Rev Genet 3: 199-209.
69. Yang A, Schweitzer R, Sun D, Kaghad M, Walker N, et al. (1999) p63 is essential for regenerative proliferation in limb, craniofacial and epithelial development. Nature 398: 714-718.
70. Hibi K, Trink B, Patturajan M, Westra WH, Caballero OL, et al. (2000) AIS is an oncogene amplified in squamous cell carcinoma. Proc Natl Acad Sci U S A 97: 5462-5467.
71. Yang A, Kaghad M, Wang Y, Gillett E, Fleming MD, et al. (1998) p63, a p53 homolog at 3q27-29, encodes multiple products with transactivating, death-inducing, and dominant-negative activities. Mol Cell 2: 305-316.
72. Westfall MD, Pietenpol JA (2004) p63: Molecular complexity in development and cancer. Carcinogenesis 25: 857-864.
73. Wu G, Nomoto S, Hoque MO, Dracheva T, Osada M, et al. (2003) DeltaNp63alpha and TAp63alpha regulate transcription of genes with distinct biological functions in cancer and development. Cancer Res 63: 2351-2357.
74. Koster MI, Roop DR (2004) The role of p63 in development and differentiation of the epidermis. J Dermatol Sci 34: 3-9.
75. King KE, Ponnamperuma RM, Yamashita T, Tokino T, Lee LA, et al. (2003) deltaNp63alpha functions as both a positive and a negative transcriptional regulator and blocks in vitro differentiation of murine keratinocytes. Oncogene 22: 3635-3644.
76. Dohn M, Zhang S, Chen X (2001) p63alpha and DeltaNp63alpha can induce cell cycle arrest and apoptosis and differentially regulate p53 target genes. Oncogene 20: 3193-3205.
77. Ihrie RA, Marques MR, Nguyen BT, Horner JS, Papazoglu C, et al. (2005) Perp is a p63-regulated gene essential for epithelial integrity. Cell 120: 843-856.
78. Kurata S, Okuyama T, Osada M, Watanabe T, Tomimori Y, et al. (2004) p51/p63 Controls subunit alpha3 of the major epidermis integrin anchoring the stem cells to the niche. J Biol Chem 279: 50069-50077.
79. Dellavalle RP, Egbert TB, Marchbank A, Su LJ, Lee LA, et al. (2001) CUSP/p63 expression in rat and human tissues. J Dermatol Sci 27: 82-87.
80. Pellegrini G, Dellambra E, Golisano O, Martinelli E, Fantozzi I, et al. (2001) p63 identifies keratinocyte stem cells. Proc Natl Acad Sci U S A 98: 3156-3161.
81. Maruya S, Kies MS, Williams M, Myers JN, Weber RS, et al. (2005) Differential expression of p63 isotypes (DeltaN and TA) in salivary gland neoplasms: biological and diagnostic implications. Hum Pathol 36: 821-827.
82. Frisch SM, Francis H (1994) Disruption of epithelial cell-matrix interactions induces apoptosis. J Cell Biol 124: 619-626.
83. Meredith JE, et. al. (1993) The extracellular matrix as a cell survival factor. Mol Biol Cell 4: 953-961.
84. Rytomaa M, Martins LM, Downward J (1999) Involvement of FADD and caspase-8 signalling in detachment-induced apoptosis. Curr Biol 9:

1043-1046.
85. Shimada A, Kato S, Enjo K, Osada M, Ikawa Y, et al. (1999) The transcriptional activities of p53 and its homologue p51/p63: similarities and differences. Cancer Res 59: 2781-2786.
86. Reginato MJ, Mills KR, Paulus JK, Lynch DK, Sgroi DC, et al. (2003) Integrins and EGFR coordinately regulate the pro-apoptotic protein Bim to prevent anoikis. Nat Cell Biol 5: 733-740.
87. Gumbiner BM (1996) Cell adhesion: the molecular basis of tissue architecture and morphogenesis. Cell 84: 345-357.
88. Nanba D, Nakanishi Y, Hieda Y (2001) Changes in adhesive properties of epithelial cells during early morphogenesis of the mammary gland. Dev Growth Differ 43: 535-544.
89. Watt FM, Hogan BL (2000) Out of Eden: stem cells and their niches. Science 287: 1427-1430.
90. Fuchs E, Tumbar T, Guasch G (2004) Socializing with the neighbors: stem cells and their niche. Cell 116: 769-778.
91. Tumbar T, Guasch G, Greco V, Blanpain C, Lowry WE, et al. (2004) Defining the epithelial stem cell niche in skin. Science 303: 359-363.
92. Ellisen LW, Ramsayer KD, Johannessen CM, Yang A, Beppu H, et al. (2002) REDD1, a developmentally regulated transcriptional target of p63 and p53, links p63 to regulation of reactive oxygen species. Mol Cell 10: 995-1005.
93. Wingender E, Chen X, al. e (2001) The TRANSFAC system on gene expression regulation. Nucleic Acids Research 29: 281-283.

Chapter 3 - p63 regulates an adhesion program and cell survival in epithelial cells

3.1 Abstract

p63 has been shown to play a critical role in the development of stratified epithelia and its derivative tissues; however little is known about the specific cellular programs that are regulated by this transcription factor. We utilized the normal breast epithelial cell line, MCF-10A, which express basal epithelial markers as model system to investigate cellular processes regulated by p63. Through transcriptional profiling and an enrichment-based category analysis, we identified genes in cellular adhesion process as significantly upregulated in p63 over expressing cells and significantly downregulated in $p63^-$ cells, suggesting a regulatory role in cell adhesion processes. Further experimental characterization of the effects of specific loss and gain of p63 function verified the computationally suggested vital role for p63 in cellular adhesion. shRNA-mediated knockdown of endogenous ΔNp63 expression in MCF-10A cells caused cell detachment.

P.S. This research was done in collaboration with Dr. Joan Brugge's Laboratory at Harvard Medical School. I performed the computational analysis of the transcriptional profile and Dr. Danielle Lynch and et. al. at Brugge Laboratory performed the extensive experimental analysis. Here, for the completeness of this report, I include

part of the experimental results that support my computational results as the experimental verification.

3.2 Introduction

The formation of specialized epithelial tissues/appendages, such as the mammary gland, is regulated by the orchestration of complex transcriptional programmes[68]. p63, a member of the p53 family of transcription factors has been shown to play a pivotal role in the development of stratified epithelia and its derivative tissues[40,69]. However, the exact signals and transcriptional events downstream of p63 have not been clearly defined.

Insight into p63 function has been gained by genetic models in which p63 expression is disrupted or overexpressed. p63 –/– mice exhibit severe abnormalities in the development of stratified squamous epithelia and its derivatives, including epidermis, mammary glands, prostate, salivary gland and other tissues[40,69]. Ectopic *p63* expression in skin is sufficient to drive crucial aspects of stratification and if unchecked, results in the induction of metaplasia[3]. Furthermore, in fibroblasts ectopic expression of p63 induces anchorage independent growth and tumour formation in nude mice[70]. In addition germ line p63 mutations in humans are associated with the ectrodactyly, ectodermal dysplasia, limb-mammary syndrome , cleft palate syndrome and other malformation syndromes[71]. Lastly there is also increasing evidence that p63 may play a role in human cancers[70,72,73], although its precise role remains to be fully clarified. Thus, in skin, and most likely other stratified epithelia, p63 plays dual roles:

Firstly, as a molecular switch for initiating epithelial stratification/cell fate determination and secondly in the maintenance of the proliferative capacity/potential of the basal cell compartment and/or stem cells in mature tissues[74]. Such findings underscore the critical nature of p63 in epithelial tissue development; however, the mechanisms involved in p63 control of development and proliferation are not well understood.

The analysis of p63 function is complicated by the complex nature of the p63 gene itself, which encodes at least six different isoforms that differ in their N-terminus due to alternate promoter usage, and in their C –terminus due to alternative RNA splicing[71][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6]. Transactivating (TA) isoforms of p63 contain an N-terminal exon that encodes a strong transactivation domain, whereas delta N-terminus (ΔN) isoforms are lacking this domain and start upstream from the common DNA binding domain. Based on their structures, it has been suggested that TA- and ΔNp63 isoforms may have opposing functions. This possibility is supported by evidence that the ΔN isoforms act as transcriptional repressors both in vitro and in animal models and can strongly oppose p53-family -mediated transactivation in reporter assays[3,71]. However ΔN isoforms of p63 and p73, another p53 family member, have also been shown to act as positive regulators of transcription[75]. Indeed, ΔNp63 isoforms contain a second activation domain within the N-terminus that can activate transcription and is capable of inducing cell cycle arrest and apoptosis when ectopically overexpressed[75,76]. Consistent with the structural similarity of the DNA binding domain of p63 and p53 ,

many p53-regulated genes are also responsive to the p63 proteins both in vitro and in vivo, although few true endogenous p63 targets have been identified. The identification of target genes is critical to fully understanding p63 function in vivo. Recent work identified the p53 target gene, Perp, as a key effector in the p63 developmental program playing an essential role in promoting stable assembly of specialized adhesive complexes critical to epithelial tissue integrity[77]. We and other groups have identified additional potential p63 targets[73,76,78] offering some insight into diverse functions of p63.

To date, p63 function of has been characterized primarily in the context of skin and relatively little is known about its role in other tissues. Genetic ablation of p63 in mice results in the complete lack of mammary gland development highlighting the importance of p63 in the development of this tissue[40,69] however in these studies, there is a distinct lack of information regarding its role/function in mammary epithelial biology. p63 expression is somewhat restricted to the proliferative basal cell layer in a wide range of epithelial tissues, including the myoepithelial/basal cells within the mammary gland[69,79]. Furthermore, the predominant isoforms expressed being the non-transcactivating ΔNp63 isoforms, particularly ΔNp63α, to the near exclusion of TA isoforms[71,72,80,81] suggesting that ΔNp63 isoforms must be playing a major role in the biology of this cell type. The function of basal/myoepithelial cells and their role in cancer and development are not well understood. They mediate the interaction between ductal luminal cells and the extracellular matrix and provide primary structural support and contractility during lactation. In addition, they have been shown to be able

to suppress breast cancer cell growth, invasion and angiogenesis and due to these properties have been described as 'natural tumour suppressors'. Cells of a basal epithelial phenotype are the earliest detected during the development of the mammary gland and possibly mark early mammary progenitor cells. A subset of highly aggressive breast cancers (15%) that have poor clinical outcome exhibit a basal epithelial phenotype and are characterized by high levels of expression of basal cytokeratins, laminin γ2 and α3 chains, Fibronectin and β4- and α6 integrin. The normal breast epithelial cell line, MCF-10A, expresses markers commonly associated with a basal/myoepepithelial phenotype, including high-molecular-weight cytokeratins and ΔNp63α expression making this a relevant model system to dissect physiological/endogenous functions of p63 on mammary epithelial biology.

In this chapter, I report the study in collaboration with Joan Brugge Laboratory at Harvard Medical School on the effects of loss, using an RNAi approach, or gain of expression of p63 in cultured human breast epithelial cells in order to examine endogenous function and biological activities regulated by p63 and used transcriptional profiling to identify downstream targets of p63 that may provide insights into its biological activities. Our computational and experimental results provide an initial understanding of the subprograms downstream of p63 and define a novel role for p63 as a critical regulator of epithelial cell adhesion.

3.3 Results

3.3.1 Loss of endogenous p63 expression induces detachment and death in mammary epithelial cells

ΔNp63α is the predominant p63 isoform expressed in epithelial tissues, however there are six major isoforms of p63 generated by alternate splicing shown in Fig. 1a and b. In order to analyze which isoforms are expressed the normal mammary epithelial cell line, MCF-10A we compared electophoretic mobility of endogenous p63 using an antibody that recognizes all p63 isoforms [71][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6]with transiently expressed constructs encoding all six isoforms. The most abundant form of p63 detected in MCF-10A cells using an antibody that recognizes all p63 isoforms is the ΔNp63alpha form (Fig.20A). [71][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6][6]This form comprises more than 90% of all detectable p63 protein in MCF10-A. Consistent with this finding, comparison of the mRNA levels of DNp63 versus TAp63 isoforms using quantitative RT-PCR (QRT-PCR) indicates that DNp63 isforms are expressed at levels more than 10-fold fold higher than TAp63 isoforms (data not shown).

To investigate the function of endogenous p63 in MCF-10A cells and to assess the relative importance of the individual p63 isoforms, we disrupted expression of subsets of p63 isoforms using adenovirus-transduced short hairpin RNA's (shRNA) Ffigure 20b). The efficiency and specificity of knockdown was monitored 48hr following adenoviral adenoviral transduction of the shRNA by western blotting and QRT-PCR (Figure

20c). Ablation of all TAp63 isoforms (TAp63α, TAp63β, and TAp63γ) had little effect on MCF-10A morphology relative to control infected cells (Figure 20d). However, ablation of alpha isoforms or all p63 isoforms using shRNA targeted against either the alpha tail or the core DNA binding domain (DBD) common to all known p63 isoforms, respectively, had pronounced phenotypic effects. Cells lacking either the alpha or all isoforms of p63 displayed a rounded morphology and detached from the plate (Figure 20d). The effect of both alpha and all p63 isoform knockdown is most likely due to down-regulation of ΔNp63α because 1) knockdown of TAp63 isoforms (α, β and γ) did not induce a similar effect, 2) ΔNp63α is the predominant isoform expressed in MCF-10A cells and 3) ΔNp63α expression is lost following transduction with the shRNA against alpha p63 isoforms.

Given that matrix adhesion is required for epithelial cell survival[82-84] we examined whether the loss of p63 induces apoptosis in addition to cell detachment. Cell death was analyzed in cells 48hrs following infection with shRNA adenoviral vectors using three methods: sub-G1 DNA content as assessed by Fluorescence Activated Cell Sorting (FACS), a DNA fragmentation Elisa assay, and western blotting for proteins cleaved by apoptotic caspases. Downregulation of all or alpha, but not TA isoforms of p63, resulted in an increase in sub-G1 DNA content, increased DNA fragmentation, and elevated caspase 3 and PARP cleavage. Together these findings imply that loss of DNp63a causes an induction of apoptosis (Figure 20e). To confirm this interpretation, and to address the specificity of these shRNA-induced effects, we evaluated whether expression of a

variant of ΔNp63α that is resistant to the shRNA species used to downregulate all or alpha isoforms could rescue these effects. Expression of an shRNA-insensitive mutant of ΔNp63α, but not a similar TAp63γ mutant, blocked cell detachment and inhibited induction of apoptosis following p63 knockdown (data not shown). Together these data indicate that loss of ΔNp63a, the major p63 isoform expressed in MCF-10A, causes cell detachment and death. Thus, ΔNp63a is essential for MCF-10A cell survival.

To address whether the cell detachment was a secondary consequence of the cell death/apoptosis resulting from p63 downregulation, we established stable pools of MCF-10A cells overexpressing the anti-apoptotic protein Bcl-2 and subjected these cells to p63 downregulation by expression of a p63 DBD shRNA that targets all p63 isoforms. Bcl2 expression was sufficient to block apoptosis induced by p63 loss (Figure 20f) but had little effect on cell detachment (Figure 20f). These results indicate cell detachment following p63 loss is independent of apoptosis, and suggests that apoptosis may be, at least in part, secondary to cell detachment, thus resembling the process referred to as anoikis (death caused by detachment from matrix).

3.3.2 p63 regulates an adhesion subprogram

To further investigate endogenous p63 function in mammary epithelial cells and to elucidate possible mechanisms by which p63 loss causes cell detachment we took a non-biased approach and used transcriptional profiling to identify alterations in gene expression following downregulation of p63. As a complementary approach, we

also analyzed the effects of ectopic p63 expression using retroviral transduction to express two isoforms of p63 in MCF-10A cells: ΔNp63α, the most abundant isoform expressed in both epithelial tissues and MCF-10A cells, and TAp63γ, which most closely resembles p53 in structure and transactivation activities[71,85]. Expression levels of ΔNp63α were analysed by immunoblotting (Figure 21a left panel), however because TAp63γ is barely detectable at the protein level (most likely due to its rapid turnover) we analysed its expression by RT-PCR (Figure 21a right panel). Both ectopically transduced isoforms were expressed at levels approximately 4-fold greater than the respective endogenous isoforms.

To compare gene expression profiles following either loss or gain of p63 function, RNA was isolated from cells 48hr following infection with adenoviral DBD or TA shRNA vectors described above, or following infection with retroviral vectors encoding either TAp63γ or ΔNp63α. (data not shown, Figure 21a and b). Changes in gene expression were analysed using Affymetrix U133 2.0 genechip arrays. At this time point expression levels were maximal for p63 gain of function; in cells transduced with the shRNA vectors p63 was downregulated greater than 90% and phenotypic effects due to alteration in p63 levels were observed (see Figure 20). Transcriptional profiles for six populations of cells (Loss of function; Control, TA si and DBD si; Gain of function: Control, ΔNp63α and TAp63γ) were obtained and analyzed to identify distinguishing features of both gain and loss of p63 function.

Using the described criteria for the definition of candidate target genes, downregulation of endogenous p63 by the DBD shRNA sequence reduced the expression of 1063 genes and upregulated 950; whereas the TA-specific shRNA downregulated 348 genes and upregulated 415. ΔNp63α expression upregulated 336 genes and downregulated 235 genes, whereas TAp63γ induced four times as many genes (1420) and downregulated 1331 genes. Since loss of p63 induced cell detachment we asked whether the regulated gene set was enriched in genes involved in cell adhesion. Based on the gene ontology cell adhesion classification, we observed a strong bias toward downregulation of the cell adhesion category by the DBD shRNA (p value = 6.75e-5), particularly in the cell-matrix adhesion group (15/224 versus 1063 /22277, total p= 5.61e-7). Consistent with these findings, expression of both ΔNp63α and TAp63γ selectively induced genes involved in cell adhesion, albeit with lower statistical confidence (p=0.022, p=0.047, respectively). This difference may reflect the increased specificity of our shRNA approach to identify true endogenous p63 target genes. Genes encoding many types of cell adhesion proteins were regulated by modulation of p63, including ECM components (e.g.laminin and collagen subunits, fibronectin), integrins (β1, β4, β5, β6, α6, α10), components of adherens and desmosomal junctions (FAT, cadherin 4, desmoglein), other adhesion receptors (DDR1, CD44, , CD47, ,), and intracellular components of adhesion complexes (zyxin, Pyk2). As expected, many genes that displayed reduced levels of expression when total p63 was reduced with the DBD shRNA showed elevated levels of expression in the context of p63 overexpression (38 genes, see clusters 2A, B, C) Conversely, genes that were upregulated when p63 was downregulated displayed reduced levels of

expression when p63 was overexpressed (9 genes, see cluster D). However, this correlation was not always observed (e.g.tenascin). Overall, our analysis indicates with high statistical confidence that endogenous p63 functions in MCF-10A to regulate preferentially genes involved in cell adhesion.

Several specific clusters of genes that displayed distinct patterns of adhesion gene regulation are shown in figure 2b. Cluster A shows genes that were downregulated by the DBD hairpin, but not the TA hairpin and were upregulated by both ΔN and TA forms of p63 (includes NCAM, collagen type VII, and the integrin associated protein CD47). This pattern of expression may represent genes that are regulated by endogenous ΔNp63α, but they can be induced by either isoform when overexpressed. Cluster B contains genes that were downregulated by both hairpin vectors and upregulated by overexpression of both forms of p63 [includes all three chains of laminin 5, integrin β2 and β5, the protocadherin FAT, the FAK-related kinase Pyk2, zyxin, nectin, ICAM-1, the protein tyrosine phosphatase PTPRF, and thrombospondin 1]. These genes may represent those for which both endogenous ΔNp63 and TAp63 isoforms contribute as positive regulators. Cluster C genes were downregulated most significantly by DBD, but upregulated most significantly by TA (includes the collagen receptor DDR1, fibronectin, two cadherins, delta catenin, collagen V and plakoglobin). Cluster D genes were upregulated by reduction in either or both hairpin vectors and downregulated by one or both p63 cDNAs (plakophilin, laminin beta2, collagen VI, dlg 5, protocadherin).

To validate some of the changes observed by transcription profiling we performed QRT-PCR on several candidate genes that were identified independently in both loss of function and gain of function experiments. We focused on genes involved in cell-matrix adhesion that were strongly influenced by alterations in p63 levels. These included the ECM components laminin γ2 (LAMC2) and fibronectin (FN), and integrins β1 (ITGB1) and β4 (ITGB4). Expression of β1, β4 and α6 integrin were downregulated following reduction of all or alpha p63 isoforms, and only slightly or not at all affected by reduction of TA p63 isoforms (Figure 21c). Conversely, ectopic expression of ΔNp63α, increased β1, β4 and α6 integrin levels, whereas exogenous TAp63γ expression caused a slight reduction in mRNA levels of all three integrin genes. Laminin γ2 and fibronectin were also reduced following p63 knockdown. Interestingly, while laminin γ2 was upregulated to a similar extent by both p63 isoforms, only TAp63γ induced fibronectin expression, providing an example of a gene that is differentially regulated by ΔNp63 and TAp63 isoforms. We also analysed several known p63 targets such as p21, Jagged-1 and Perp by QRT-PCR and confirmed them as p63 targets in MCF-10A cells (data not shown).

3.3.3 Regulation of cell adhesion proteins by p63

In order to determine if alterations observed at the mRNA levels were translated to changes in adhesion protein levels we expressed DBD, TA or control shRNA in MCF-10A cells using adenoviral infection and analysed lysates by western blotting 48h following infection. We found that complete p63 ablation caused a marked reduction or loss of β1-, and β4- integrins and slight reduction in α6 integrin at the protein level

(Figure 22a). We additionally looked at EGFR levels, since a previous study suggested that EGFR is a transcriptional target of p63, and because we have previously shown that EGFR expression is lost following cell detachment of MCF-10A cells[86]. Downregulation of all p63 isoforms caused a marked reduction in EGFR levels (Figure 22a). However downregulation of TA isoforms had little or no effect on expression of any of these proteins, suggesting that ΔNp63 isoforms are the major isoforms controlling maintenance/expression of cell adhesion molecules. Importantly, the reduction of β1-, β4 integrins or EGFR levels caused by p63 loss was not affected by Bcl2 expression (Figure 22b) suggesting that these events are independent of cell death. Furthermore, shRNA-mediated knockdown of alpha p63 isoforms caused a reduction in β-integrins identical to that observed with complete p63 ablation using the DBD shRNA to (data not shown) suggestive that ΔNp63α is essential for cell adhesion processes.

To determine if there are complementary changes in protein expression following ectopic expression of p63 we analysed cells 48h following retroviral transduction with either ΔNp63α, TAp63γ or vector control. β1 integrin was elevated only in cells expressing ΔNp63α and not TAp63γ – indeed, there was a slight reduction in β1 levels in TAp63γ expressing cells relative to control, which is consistent with changes seen at the mRNA level (Figure 22c). Levels of α6 integrin were elevated to the same extent (approx. 3-fold relative to control cells) by both isoforms of p63. We additionally examined MCF7 cells, which have much lower basal levels of many of the integrin subunits and undetectable

levels of ΔNp63α. Ectopic expression of ΔNp63α upregulated integrins β1, β4 and α6 at least 2 fold (data not shown).

Since endogenous levels of ECM components were low or undetectable in parental cells, the loss of expression was not detectable at a protein level following shRNA mediated p63 reduction (data not shown). Therefore, we also analyzed expression of several ECM components (fibronectin, laminin I and V) following p63 expression in MCF-10A cells. All of these proteins were upregulated in cells expressing either isoform of p63 although to a much greater extent in the TAγ expressing cells (Figure 22d). Furthermore entactin/nidogen was upregulated only in cells expressing TAp63γ (Figure 22d) confirming the specific increase in mRNA expression observed in the microarray analysis (supplemental Table 2). These data strongly support a role for p63 in the regulation of cell adhesion programs, particularly those involving integrin and ECM components.

3.3.4 Cell adhesion is regulated by p63 levels

Attachment to extracellular matrix and spreading are mediated by integrins and other matrix receptors. Engagement of integrins leads to tyrosine phosphorylation of a well chararcterised set of proteins including Cas (Mr130K) FAK/Pyk (Mr115-120K) and paxillin (Mr65-70K), which link directly or indirectly with integrins. Since alterations in p63 levels markedly changed integrin/ECM protein and gene expression levels we examined whether elevated expression of p63 affected the profile of tyrosine phosphorylated proteins. Lysates from MCF-10A cells infected with retroviruses

encoding either ΔNp63α or Tap63γ for 48h were immunoblotted with a phosphotyrosine-specific antibody. Relative to control cells, those expressing either isoform of p63 displayed an increase in cellular tyrosine phosphorylation of proteins with electrophoretic mobilities similar to those of Cas, FAK and paxillin (Figure 23a). Furthermore we confirmed that FAK, Pyk2 and paxillin were highly tyrosine phosphorylated in cells expressing either isoform of p63 using phospho-specific antibodies to each protein (Figure 23a). While the total levels of FAK and paxillin were not affected by p63 expression the expression of Pyk/FAK2 was increased several fold (Figure 23a). These data indicate that ectopic expression of p63 can alter integrin-mediated cell adhesion signaling, supporting the notion that p63 regulates cell adhesion.

To assess the functional consequences of alterations in p63 expression on cellular processes regulated by adhesion receptors, we examined cell adhesion in cells expressing elevated or decreased levels of p63. Firstly we assessed the ability of cells stably expressing either p63 cDNAs or p63 shRNAs to adhere to a variety of exogenous matrix proteins (laminin I, basement membrane complex (BMC), fibronectin and collagen IV). Increased expression of both TAγ and ΔNα isoforms of p63 enhanced adhesion to laminin I (2 and 5 fold respectively), BMC (2.6 and 3.6 fold), fibronectin (19 and 17 fold) and collagen (2 and 3 fold) relative to control cells (Figure 23b). The fold increase in adhesion to fibronectin was much more pronounced than to other matrix proteins because there is relatively little binding of parental or control cells to this substrate. The ability of ΔNp63α to confer enhanced adhesion to all matrix types relative to TAγ is

consistent with our model that ΔNp63α is the major p63 isoform regulating the cellular adhesion program in MCF-10A.

To examine the effects of loss of p63 on matrix adhesion we examined cell attachment to matrix proteins at 24h following p63 downregulation, since the majority of cells lacking p63 were detached or very loosely adherent by 48hrs At the 24h time point, greater than 50% of cells lacking either all or alpha p63 isoforms were still adherent (data not shown). Loss of all or alpha isoforms caused a pronounced diminution in cell adhesion to all substrates (Figure. 23c). These observations support the findings of our overexpression studies, and they indicate that p63, in particular alpha isoforms of p63, are required for matrix adhesion Interestingly, TA isoform-specific downregulation caused a significant increase in the ability of MCF-10A cells, to adhere to exogenous laminin I but not to the other substrates, suggesting that TAp63 isoforms may oppose the function of ΔNp63α and negatively regulate binding to laminin I. Similar results were obtained for cells infected with the shRNA adenovirus at an earlier time point 12hr (data not shown). Furthermore, the reduction in adhesion to exogenous matrix following p63 knockdown was unaffected by stable Bcl2 expression, suggesting that functional loss of adhesion to exogenous matrix precedes cell death (Figure 23d).

3.4 Discussion

p63 has been shown to play a critical role in the development of the mammary gland and other stratified epithelia; however there is little known about the specific cellular

programs that are regulated by this transcription factor. In this report we utlised the normal breast epithelial cell line, MCF-10A, which expresses p63 and other basal epithelial markers as an in vitro model system to investigate cellular processes that are regulated by p63. We used a combined analysis of both loss and gain of p63 function to identify a relevant transcriptional program. As expected, the effects of loss and gain of p63 expression did not show a reciprocal regulation for all genes. For example, while many genes, that were downregulated by the DBD siRNA were upregulated by ectopic expression of one or both isoforms of p63, such as the integrins noted above, some genes that were downregulated by the DBD siRNA were not significantly upregulated by p63 overexpression. Many of the genes that showed the latter pattern are expressed at high levels in MCF-10A cells, thus obscuring detection effects of an increase in expression of p63. For example, we were able to detect an increase in β4 integrin protein expression in MCF-7 cells upon ectopic p63 expression, but this was barely detectable in MCF-10A cells. Alternatively, the lack of correspondence may reflect functional difference between the isoforms, or artifacts driven by overexpression or may reflect direct or indirect gene targets. Within the timescale used in these studies it is not clear whether the cell adhesion targets are direct or indirect, 48hr following retroviral infection would allow both direct and indirect targets to be transcribed but the effect on cell adhesion and alteration in expression of cell adhesion related genes is clear. Interestingly amongst the regulated genes many contain p53 response elements either within their promoter sequences or within the first intron raising the possibility many of these could be direct p63 targets (e.g laminin γ2).

Epithelial development and the formation of squamous epithelial derived tissues are complex processes involving ectoderm-mesenchyme cross talk and multiple secreted factors as well as cell-cell and cell-matrix interactions [68].Regulation of cell adhesion is a general feature underlying early morphogenesis of several ectoderm-derived organs including the mammary gland[87,88]. Commitment of specialized progenitor or stem cells requires extensive signaling and interactions with non stem cells and basal lamina within a specialized niche[89,90]. Adhesion proteins such as cadherins/catenins via adherens junctions and integrins via interactions with the extracellular matrix are thought to play a major role in stem cell biology within these specialized microenvironments. Indeed, loss of function studies in mice have revealed that both integrins and adherens junctions play critical roles in maintaining the location, adhesiveness and the proliferative status of epithelial stem cells within tissues (reviewed by[89]). Transcriptional profiling of these specialized cells has highlighted the importance of integrins, their ECM ligands, cell-adhesion and polarity proteins, furthermore increased levels of expression of integrins are often characteristic of stem cells[91]. Loss and/or alterations in integrin expression allows departure from the stem cell niche through differentiation or apoptosis, modulation of basement membrane composition and the local concentration of secreted factors available within the stem cell niche[89]. Given that p63 can regulate many of these cell adhesion associated genes implicated in stem cell biology it is tempting to speculate that p63 may play a major role in stem/progenitor cell biology and the regulation of adhesion involved in epithelial morphogenesis. In the mammary

gland the myopepithelial cells, which express high levels of p63, are the earliest detected during the development of the mammary gland and possibly mark early mammary progenitor cells. This cell type mediates the interaction between ductal luminal cells and the secreted extracellular matrix providing primary structural support and contractility during lactation and is characterized by their high levels of expression of integrins and ECM proteins not seen within the luminal cell population, further supporting a fundamental role for p63 in the biology of these cells.

The absence of p63 inactivation in human cancers has ruled out a typical tumour suppressor role. Instead, ΔNp63α is hypothesized to play an oncogenic role in several epithelial cancers, most notably squamous cell carcinomas, in which increased p63 expression is often accompanied by genomic amplification[70]. Indeed, ΔNp63α has been implicated in cell proliferation potential and oncogenic growth[70]. A recent comprehensive analysis of p63 expression in normal and neoplastic tissue showed that p63 expression was rarely detected in adenocarcinomas of the breast, lung and prostate, all of which lack basal epithelial cells, consistent with restricted p63 expression in squamous or basal epithelium. In keeping with this finding, p63 is expressed selectively in a subset of highly aggressive breast cancers (15%) that exhibit a basal/myoepithelial phenotype and have a poor clinical outcome. Interestingly we find that alterations in p63 expression can influence expression of many of the genes characteristic of this tumour type. These include cell adhesion proteins and ECM components such as laminin γ2 and α3 chains, fibronectin and β4- and α6 integrin, as well as EGFR. It is tempting to

speculate that p63 may contribute to oncogenesis in this type of breast cancer, but further investigation and a better understanding of p63 function and regulated target genes is needed.

In conclusion, we have shown that p63 is critical for basal epithelial cell adhesion and survival and that this regulation is mediated by transcription of a cell adhesion subprogram. The precise mechanisms by which p63 exerts these functions remain poorly defined and are the focus of current investigations.

3.5 Experimental procedures

3.5.1 Cell culture and treatments (from Danielle Lynch at Brugge Laboratory)

MCF10-A cells were maintained as described in [86]. Primary human mammary epithelial cells (HMEC) obtained from Clonectics (Cambrex) were maintained in MEGM supplemented with bovine pituitary extract. Primary human epidermal keratinocytes (HFK) were cultured in keratinocyte serum-free medium (GIBCO-BRL) containing EGF, bovine pituitary extract (BPE)) and 0.4mM Ca as described[92]. 293T cells were maintained in DMEM with 10% v/v FCS. Primary mouse mammary epithelial cells (MMEC) were obtained from Balb/C, p63 floxed mice (p63fl/fl) or wild type (WT) littermates [40] and maintained as previously described[86]. Generation of VSV-G pseudotyped retrovirus and retroviral infection of MCF10A cells was carried out

as described[11]. To determine effect of p63 isoform expression on cell growth stably infected MCF10A cells (4000/well 24 well plate) were plated and grown in assay media[11] in the absence of EGF, cells were counted (triplicate wells/timepoint) on days 2, 4, 6, 8 and 10 after plating. Cell death was measured by Propidium Iodide staining followed by flow cytometric analysis or using the cell death detection ELISA kit (Roche Diagnostics, Mannheim, Germany) according to the manufacturer's instructions each experiment was performed, at least, in triplicate.

3.5.2 Reagents, Antibodies and DNA constructs

Commercial antibodies for immunoblotting were obtained from the following sources: Integrin β1(clone 18), Integrin β4(7), EGFR(18), FAK (77), paxillin (349), Pyk2 (11) and phospho-Tyrosine(4G10) (BD Biosciences, San Jose, CA USA), p63 (4A4); Fibrononectin (IST9), β-actin, and β-tubulin (Abcam, MA USA); Integrin α6 (GoH3), nidogen/entactin and Laminin 5 (D4B5), (Chemicon, CA USA); Laminin 1 (Sigma); Collagen IV and I (Calbiochem); phospho-FAK, phosphor-Paxillin (Biosource International); cleaved Parp, cleaved caspase 3, Erk, phospho-Erk, PKB, phosphor-PKB (Cell signaling, MA USA); Bcl2, p73 and p53 (Santa Cruz, CA USA).

Human TAp63γ and ΔNp63α cDNAs and shRNA rescue mutants were subcloned as BamHI-XhoI fragments into the retroviral vector pBabe puro. shRNA rescue mutants were constructed by introducing 3 or 4 silent nucleotide changes using site

directed mutagenesis on human TAp63γ and ΔNp63α cDNAs in pcDNA3, correct incorporation of mutations was confirmed by DNA sequence analysis.

Adenoviral infection and Gene Silencing with small hairpin RNAs. MCF7 cells were infected with Ad-pShuttle-CMV-ΔNp62α or Ad-pShuttle-CMV (Ellisen 2002) for 2 hr. Cells were harvested for both protein and RNA at 24 and 48 hr after infection. p63 gene ablation was performed in vitro by cre recombinase mediated excision of floxed p63 alleles in primary MMECs. p63fl/fl and WT littermate MMECs were plated for 24 hr following isolation. Cells were trypsinised and allowed to adhere, cells were then infected with Ad5- CMV-Cre-GFP or Ad5 –CMVeGFP (Vector Development Lab Baylor College of Medicine) for 2 hr. Cells were harvested for both protein and RNA at 24 and 48 hr after infection. Isoform specific gene silencing was achieved by adenoviral-mediated expression of small hairpin RNAs (shRNA). Cells were grown in full medium and infected with Ad-pShuttle-U6-TA, Ad-pShuttle-U6-DBD, Ad-pShuttle-U6-Alpha or Ad-pShuttle-U6 for 2 hr. Cells were harvested for FACS, cell death Elisa, protein and RNA at 48hr following infection. Sh RNA target sequences were as follows: p63 TA isoform specific: 5'-gggattttctggaacagcctat-3'; DBD/All p63 isoform specific: 5'-gggaacagccatgcccagtatg-3'; Alpha p63 isoform specific: 5'-gggtgagcgtgttattgatgct-3'.

RNA interference. MCF-10A cells were plated onto 6-well plates at 300,000 cells per well. After 24 h, cells were transfected with double-stranded RNA–DNA hybrids at a final concentration of 200nM annealed oligo using Oligofectamine (Invitrogen)

according to manufacturer's instructions. After 24 h of transfection, cells were placed in anoikis and cell adhesion assays or analysed by SDS–PAGE and western blotting. Oligonucleotides were obtained from Dharmacon Research (CO, USA). Control sense 5'-(GGCUGUAACUUACGUGUACUU)d(TT)-3'; control antisense, 5'-(AAGUACACGUAAGUUACAGCC)d(TT)-3'; β4-integrin sense 5'-GAGCUGCACGGAGUGUGUCdTdT-3'; β4-integrin, antisense 5'-GACACACUCCGUGCAGCUCdTdT-3' β1-integrin sense 5'-GGAUUACUUCGGACUUCAGdTdT-3' β1-integrin antisense 5'-CUGAAGUCCGAAGUAAUCCdTdT-3'

3.5.3 Microarray Analysis

Total RNA was isolated by phenol-chloroform extraction (TriReagent, Sigma) 48 hr following retroviral or adenoviral infection and was subjected to reverse transcription, labeling and hybridisation to U133 v2.0 gene chip arrays (Affymetrix, CA, USA) containing 14,500 human genes, each experiment was performed in triplicate. Background correction and normalization was done with MAS5 function in Bioconductor Affy package{Laurent Gautier, Leslie Cope, Benjamin Milo Bolstad, and Rafael A. Irizarry. Affy – an r package for the analysis of affymetrix genechip data at the probe level. Bioinformatics, 2003.} Quality control was done using Bioconductor AffyPLM package. The differential expression was assessed using the empirical Bayes method implemented in Bioconductor Limma package, and the p values were adjusted by false discovery rates. The significance level was control at false discovery rate 0.05, and

only genes whose induction or loss was verified by RT-PCR or western blotting were analyzed further.

3.5.4 Enrichment analysis

We assessed differentially expressed genes for the significant enrichment of Gene Ontology Cell adhesion category and all its subcategories. The enrichment analysis computes the probability that the number of genes in a specific type of regulation (e.g down regulated in DBDsi) being annotated as within a specific cell adhesion category (e.g. "cell-matrix adhesion") would occur by chance, and the results are given in the form of p values.

3.5.5 Protein preparation and Immunoblotting (From Brugge laboratory)

Protein Lysates were prepared as previously described using modified RIPA buffer. Protein concentration was determined using a Bradford dye-based assay (Biorad). 20µg total protein was subjected to SDS-PAGE followed by immunoblotting with appropriate antibodies at recommended dilutions followed by incubation with peroxidase linked secondary antibodies and enhanced-chemiluminescent detection.

3.5.6 RNA isolation and RT-PCR (From Brugge laboratory)

Total RNA was isolated from cells grown in 10cm dishes using Tri-Reagent (Sigma) according to manufacturers instructions. 0.5µg total DNAse-treated RNA was then

amplified using gene specific primers with the One-Step RT-PCR kit (Qiagen) according to manufacturers instructions. All PCR products were analysed by gel electrophoresis. Quantitative RT-PCR was performed using Quantitect SYBR Green RT-PCR kit (Qiagen) according to manufacturers instructions. Primer sequences are available on request. All primers spanned at least one intron and control amplification was performed on RNA samples not subjected to the reverse transcription in parallel to ensure no contaminating genomic DNA was present.

3.5.7 Cell-adhesion assays (From Brugge laboratory)

Cell adhesion to laminin I, collagen IV, fibronectin or basement membrane complex was performed using the Innocyte(™) ECM cell adhesion assay (Calbiochem,) according to manufacturer's instructions. Briefly 10,000 cells were allowed to adhere to exogenous matrix for 1h 37°C in a 96-well plate. Wells were subsequently washed thoroughly to remove non-adherant cells and calcein-AM dye added to allow quantification of adherent cells. Following 1h incubation with calcein AM at 37°C fluorescence was measured at 485/520nm.

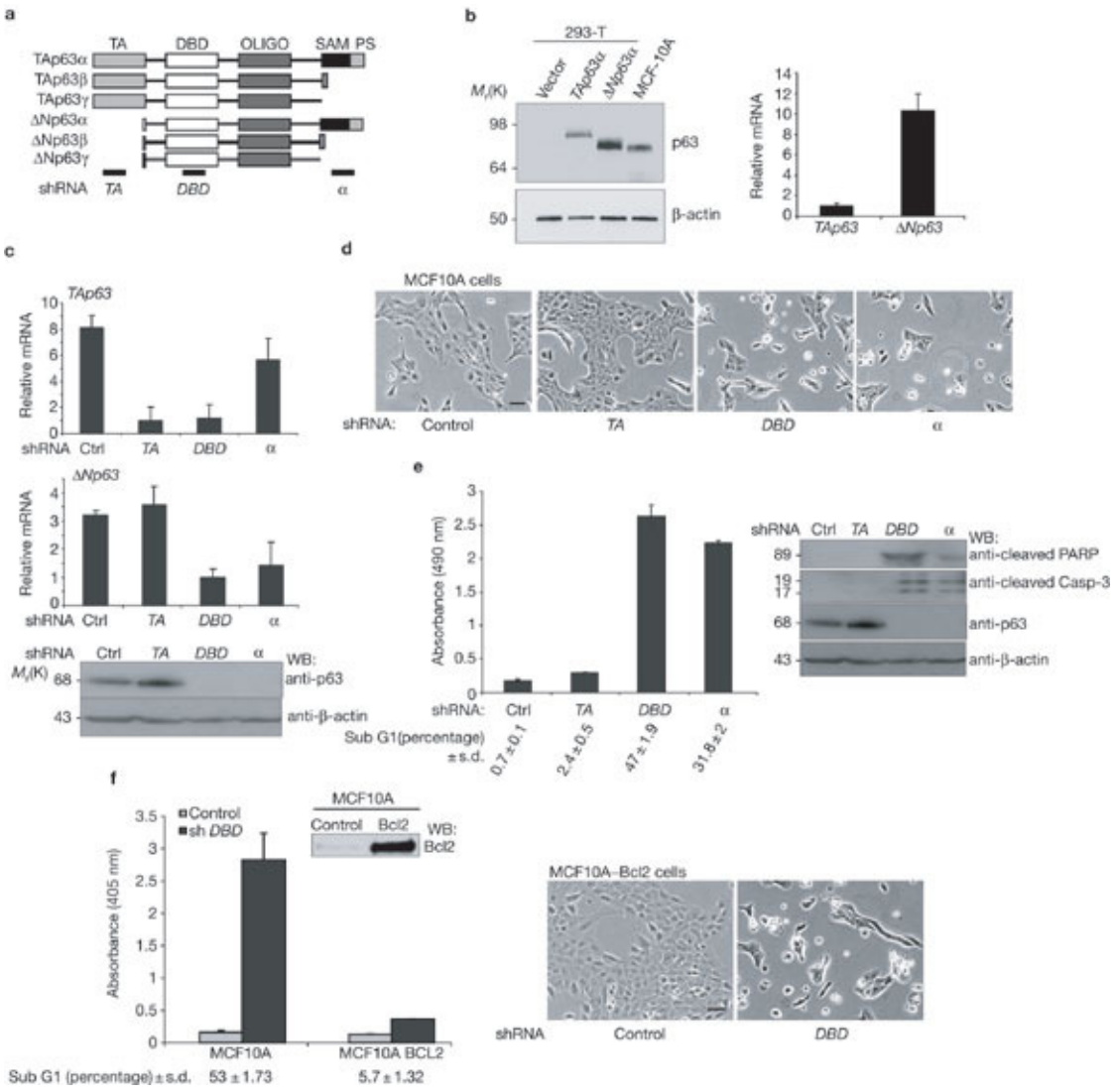


Figure 20. Loss of endogenous p63 expression induces detachment and death in mammary epithelial cells. (a) Expression levels of p63 isoforms in MCF10A cells was determined by western blotting and compared with mobility of the six major p63 isoforms transiently expressed in 293T cells. (b) Schematic representation of p63 isoforms and relative position of shRNA sequences. TA; (transactivation domain) DBD; DNA binding domain, Oligo; oligomerization domain, SAM; Sterile alpha motif domain, and PS; post-SAM domain. (c) Expression levels of p63 isoforms in MCF10A cells following isoform specific knockdown using adenovirally transduced shRNA 48h following infection. Expression of ΔNp63α was determined by western blotting, expression of TA isoforms was assessed by qRT-PCR shown graphically, values represent the mean and standard deviation of three independent experiments. (d) Effects of p63 isoform specific downregulation on cellular morphology. Phase contrast micrographs show morphology of MCF10A cells 48hr following infection with adenoviral vectors encoding control or p63 isoform specific shRNAs (TA: TA

specific shRNA sequence targets α, β, and γ TA isoforms, DBD: targets the core DNA binding domain present in all p63 isoforms, Alpha: α-isoform specific shRNA targets both ΔNp63α and Tap63α isoforms). (e) Loss of p63 causes detachment induced cell death. Cells were harvested 48hr following infection with control or p63 shRNA's and were assayed for apoptosis by both cell death Elisa and FACs analysis (left panel). Values represent the mean and standard deviation of three independent experiments. Cell lysates were analysed for proteins indicative of apoptosis by western blot analysis (right panel). (f) p63 downregulation causes cell detachment independent of cell death. MCF10A cells stably expressing Bcl2 were subjected to p63 knockdown by shRNA encoding adenovirus as described in 1D. Bcl2 expression protects from cell death induced by p63 loss (upper panel). Cells were analyzed 48hrs later for cell death by cell death Elisa and FACs analysis. Values represent the mean and standard deviation of three independent experiments. Phase contrast micrographs show morphology of shRNA adenoviral infected MCF10A/Bcl2 cells 48hr following infection with control or p63 DBD (lower panel).

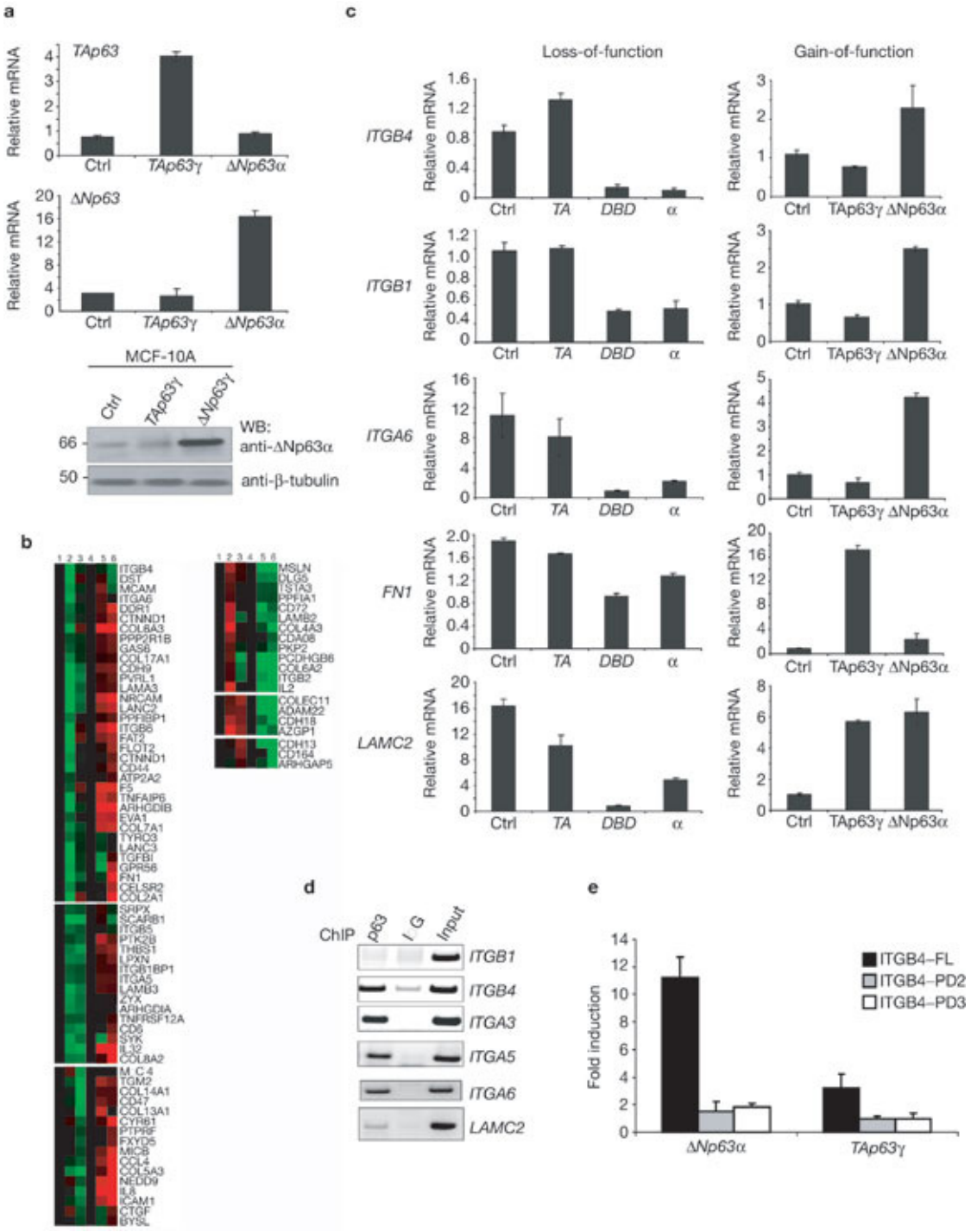


Figure 21. Identification of an adhesion subprogram regulated by p63. (a) Expression levels of p63 isoforms (TAγ: Tap63γ and ΔNα: ΔNp63α) relative to vector control (Ctrl) infected cells determined by western (left panel) and RT-PCR (right panel) (b) Microarray analysis of genes involved in cell adhesion following gain or loss of p63 function. Heat maps of gene changes greater than 2 fold (P=0.01) induced by

exogenous expression of either ΔNp63α or TAp63γ (Gain) or loss of TA or all p63 isoforms (loss). (c) Validation of microarray data: Quantitative RT-PCR analysis on RNA from MCF10A cells 48h following adenoviral infection with shRNAs against specified p63 isoforms (i) or following infection with retroviruses encoding TAp63γ, ΔNp63α or vector control, (ii). Several gene targets were selected for validation including β1 integrin (ITGB1), β4-integrin (ITGB4), α4-integrin (ITGA6), Fibronectin (FN1), laminin γ2 (LAMC2) and TA or ΔN p63 isoforms. Values represent the mean and standard deviation of three independent experiments. All of these genes confirmed the initial cDNA microarray data.

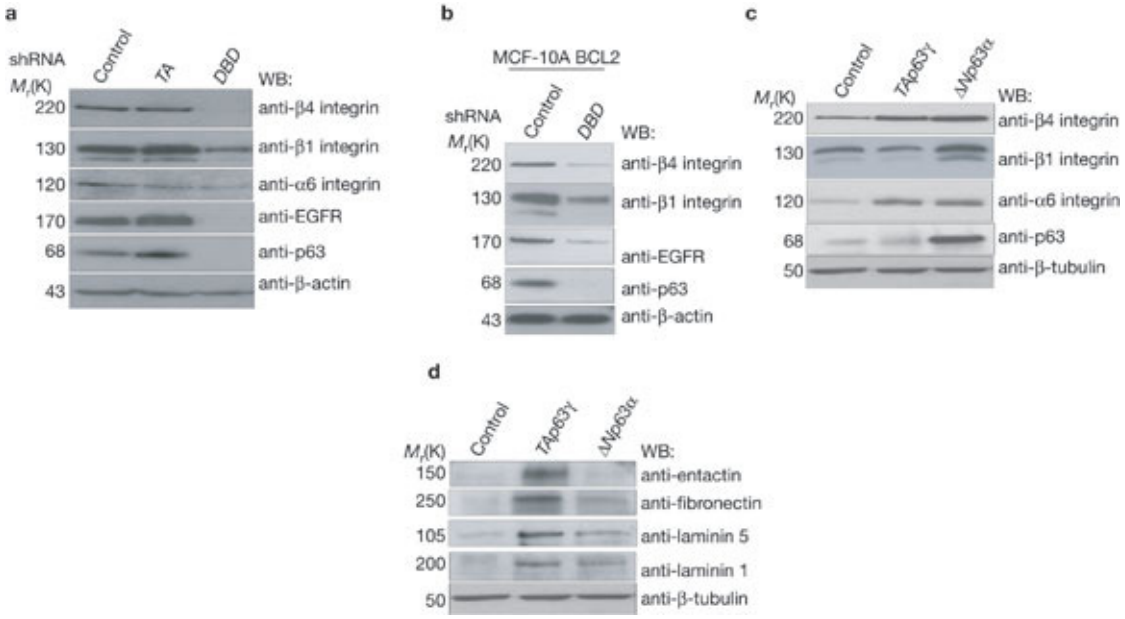


Figure 22. Regulation of cellular adhesion factors by p63. (a) Loss of ΔN but not TA p63 isoforms causes a marked reduction in cell adhesion proteins. Lysates from MCF10A cells transduced with isoform-specific p63 shRNAs expressing or control adenovirus, were anaylysed by western blotting with the indicated antibodies 48hrs following infection. (b) Reduction of cellular adhesion proteins mediated by p63 loss is independent of cell death. Lysates from cells stably expressing Bcl2 infected with p63 DBD shRNA expressing or control adenovirus, were anaylysed 48hrs following infection by western blotting with the indicated antibodies. (c) Elevated p63 expression increases integrin expression levels. Cell lysates from MCF10A cells 48h following infection with virus encoding either TAp63γ or ΔNp63α isoforms or vector control were analysed by western blotted with the indicated antibodies. (d) p63 augments cellular levels of ECM components in MCF-10A cells determined by western blotting with indicated antibodies 48h following transduction with virus encoding either TAp63γ or ΔNp63α isoforms relative to control.

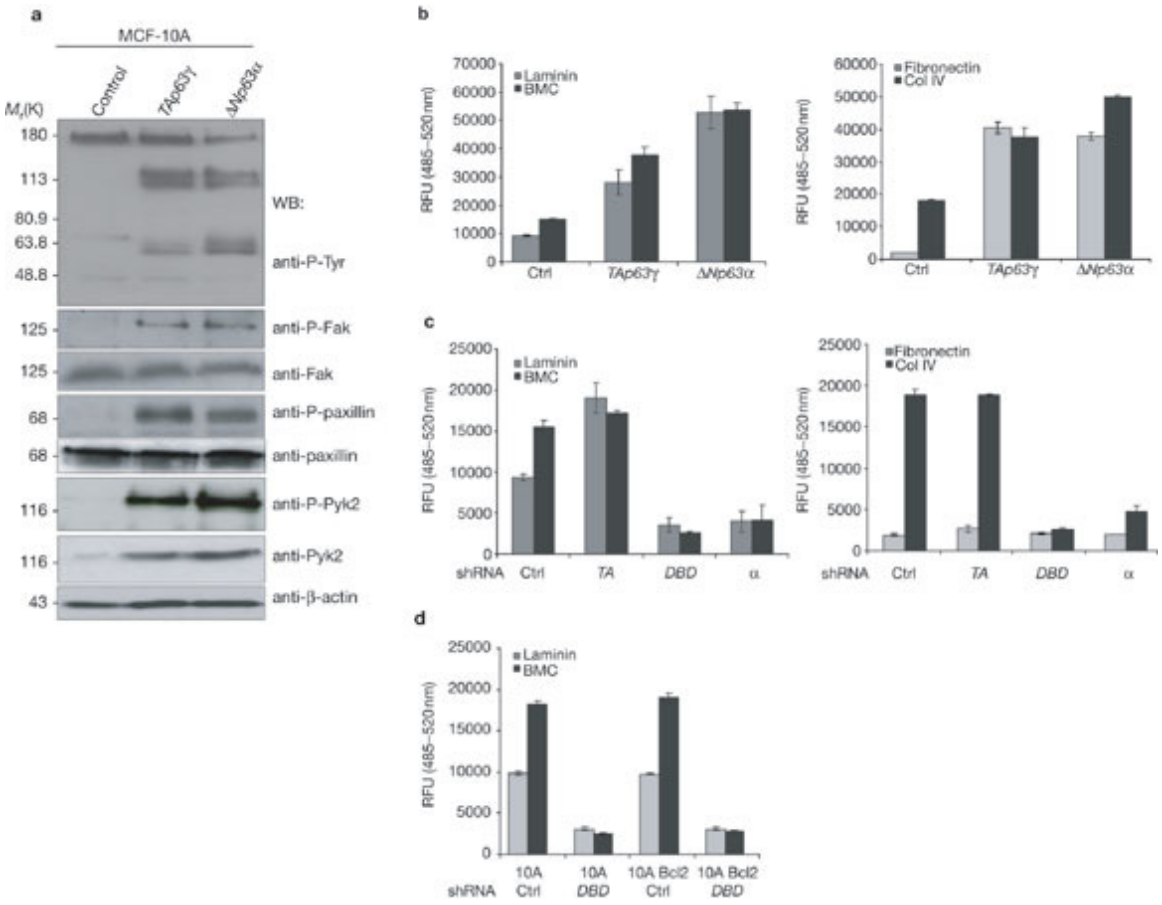


Figure 23. p63 activates adhesion–integrin signalling and promotes cell adhesion. (a) p63 expression enhances phosphorylation of integrin-regulated focal adhesion proteins. MCF-10A cells infected with control, TAp63γ or ΔNp63α retroviruses were lysed 48 h after infection and analysed by western blot with indicated antibodies. (b–d) Effect of loss or gain of p63 on adhesion to basement membrane proteins. Cells were infected with viral vectors and after the indicated time were plated on dishes coated with the indicated basement membrane proteins for 1 h and then adherent cells were quantified as described in Methods. Col IV, collagen IV. Values represent the mean ±s.d. of three replicate samples from one representative experiment (n = 3). Adhesion was measured 48 h after infection with control or p63 isoform-encoding retroviruses (b). Adhesion was measured 24 h after infection with control or p63 isoform-specific shRNAs (c). Adhesion was monitored 48h following infection of control or Bcl2 expressing cells with control or p63 DBD shRNA encoding adenoviruses (d).

3.6 References

1. Fuchs E, Raghavan S (2002) Getting under the skin of epidermal morphogenesis. Nat Rev Genet 3: 199-209.
2. Mills AA, Zheng B, Wang XJ, Vogel H, Roop DR, et al. (1999) p63 is a p53 homologue required for limb and epidermal morphogenesis. Nature 398: 708-713.
3. Yang A, Schweitzer R, Sun D, Kaghad M, Walker N, et al. (1999) p63 is essential for regenerative proliferation in limb, craniofacial and epithelial development. Nature 398: 714-718.
4. Koster MI, Kim S, Mills AA, DeMayo FJ, Roop DR (2004) p63 is the molecular switch for initiation of an epithelial stratification program. Genes Dev 18: 126-131.
5. Hibi K, Trink B, Patturajan M, Westra WH, Caballero OL, et al. (2000) AIS is an oncogene amplified in squamous cell carcinoma. Proc Natl Acad Sci U S A 97: 5462-5467.
6. Yang A, Kaghad M, Wang Y, Gillett E, Fleming MD, et al. (1998) p63, a p53 homolog at 3q27-29, encodes multiple products with transactivating, death-inducing, and dominant-negative activities. Mol Cell 2: 305-316.
7. Westfall MD, Pietenpol JA (2004) p63: Molecular complexity in development and cancer. Carcinogenesis 25: 857-864.
8. Wu G, Nomoto S, Hoque MO, Dracheva T, Osada M, et al. (2003) DeltaNp63alpha and TAp63alpha regulate transcription of genes with distinct biological functions in cancer and development. Cancer Res 63: 2351-2357.
9. Koster MI, Roop DR (2004) The role of p63 in development and differentiation of the epidermis. J Dermatol Sci 34: 3-9.
10. King KE, Ponnamperuma RM, Yamashita T, Tokino T, Lee LA, et al. (2003) deltaNp63alpha functions as both a positive and a negative transcriptional regulator and blocks in vitro differentiation of murine keratinocytes. Oncogene 22: 3635-3644.
11. Dohn M, Zhang S, Chen X (2001) p63alpha and DeltaNp63alpha can induce cell cycle arrest and apoptosis and differentially regulate p53 target genes. Oncogene 20: 3193-3205.
12. Ihrie RA, Marques MR, Nguyen BT, Horner JS, Papazoglu C, et al. (2005) Perp is a p63-regulated gene essential for epithelial integrity. Cell 120: 843-856.
13. Kurata S, Okuyama T, Osada M, Watanabe T, Tomimori Y, et al. (2004) p51/p63 Controls subunit alpha3 of the major epidermis integrin anchoring the stem cells to the niche. J Biol Chem 279: 50069-50077.
14. Dellavalle RP, Egbert TB, Marchbank A, Su LJ, Lee LA, et al. (2001) CUSP/p63 expression in rat and human tissues. J Dermatol Sci 27: 82-87.
15. Pellegrini G, Dellambra E, Golisano O, Martinelli E, Fantozzi I, et al. (2001) p63 identifies keratinocyte stem cells. Proc Natl Acad Sci U S A 98: 3156-3161.
16. Maruya S, Kies MS, Williams M, Myers JN, Weber RS, et al. (2005) Differential expression of p63 isotypes (DeltaN and TA) in salivary

gland neoplasms: biological and diagnostic implications. Hum Pathol 36: 821-827.
17. Frisch SM, Francis H (1994) Disruption of epithelial cell-matrix interactions induces apoptosis. J Cell Biol 124: 619-626.
18. Meredith JE, et. al. (1993) The extracellular matrix as a cell survival factor. Mol Biol Cell 4: 953-961.
19. Rytomaa M, Martins LM, Downward J (1999) Involvement of FADD and caspase-8 signalling in detachment-induced apoptosis. Curr Biol 9: 1043-1046.
20. Shimada A, Kato S, Enjo K, Osada M, Ikawa Y, et al. (1999) The transcriptional activities of p53 and its homologue p51/p63: similarities and differences. Cancer Res 59: 2781-2786.
21. Reginato MJ, Mills KR, Paulus JK, Lynch DK, Sgroi DC, et al. (2003) Integrins and EGFR coordinately regulate the pro-apoptotic protein Bim to prevent anoikis. Nat Cell Biol 5: 733-740.
22. Gumbiner BM (1996) Cell adhesion: the molecular basis of tissue architecture and morphogenesis. Cell 84: 345-357.
23. Nanba D, Nakanishi Y, Hieda Y (2001) Changes in adhesive properties of epithelial cells during early morphogenesis of the mammary gland. Dev Growth Differ 43: 535-544.
24. Watt FM, Hogan BL (2000) Out of Eden: stem cells and their niches. Science 287: 1427-1430.
25. Fuchs E, Tumbar T, Guasch G (2004) Socializing with the neighbors: stem cells and their niche. Cell 116: 769-778.
26. Tumbar T, Guasch G, Greco V, Blanpain C, Lowry WE, et al. (2004) Defining the epithelial stem cell niche in skin. Science 303: 359-363.
27. Ellisen LW, Ramsayer KD, Johannessen CM, Yang A, Beppu H, et al. (2002) REDD1, a developmentally regulated transcriptional target of p63 and p53, links p63 to regulation of reactive oxygen species. Mol Cell 10: 995-1005.
28. Debnath J, Muthuswamy SK, Brugge JS (2003) Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. Methods 30: 256-268.

Chapter 4 – Intelligent choice of controls for ab initio motif finding

4.1 Introduction

The interaction between trans factors and cis-regulatory motifs is a common mechanism for gene regulation (at transcriptional level and post-transcriptional level). The identification of cis-regulatory motifs not only can provide valuable clue to understanding to gene regulation, to predicting the regulatory status in an unstudied cellular context, but also can provide excellent targets for genetic manipulation.

Aiming at an effective approach for ab initio motif finding, many computational approaches have been developed, and they all fall into the same computational scheme, which is looking for statistically enriched features that are represented by certain scoring function in the sample group, in comparison to a control group. Although tremendous effect has been made on the derivation of different scoring functions, the importance of choosing sample and control groups has been under appreciated. In this Chapter, I focus on this aspect of the motif finding problem, and use a case study derived from the morphogenesis study presented in the Chapter 1 as an example to show how the intelligent choice of the control groups could be a critical factor to a motif finding algorithm's performance. This Chapter also provided detailed discussion of the biological significance of the identified transcriptional motifs during the mammary acinar *in vitro* developmental process.

4.2 A case study -- Transcriptional regulation of MixPEA-identified biological processes

A core step in the process of understanding transcriptional regulation is to identify the underlying regulators, who themselves may not be transcriptionally regulated. Since MixPEA results are essentially co-functioning and co-regulated (CFCR) gene sets, they could serve as starting sets for identifying involving transcriptional factors (TFs); the underlying transcriptional regulator could be suggested by a computational approach that compares the promoter regions of genes within these CFCR groups with promoter regions of the same number of randomly sampled genes (the negative control group); a potential functioning TF would show a significantly higher enrichment of its binding sites (cis-regulatory sites) in the CFCR group than in the negative control group.

Taking advantage of the known vertebrate TF binding site collection from TRANSFAC® database[93], we generated a binding site landscape for the 1k nt upstream regions of all the genes in the MixPEA identified gene groups, using the MATCH® program[52] provided by TRANSFAC®. To reduce the false positive, we applied an algorithm for dynamic selection of MATCH score cutoff, which minimizes the likelihood of finding the same or higher copy number of one particular binding site among protein coding regions than in promoter regions of genes in a CFCR group (Figure 29). As a true cis-regulatory site (motif) would have low frequency of appearing in protein coding

regions, using coding sequences to construct the null distribution of MATCH scores naturally enforced a more stringent cutoff for "simple" motifs that appear frequently in the genome for reason other than transcriptional regulation (Figure 29). Furthermore, since the coregulation was specifically defined for the acinar development, the chosen cutoff would favor a true motif that functions in the morphogenetic process of interest (Figure 29). Once a cutoff score was chosen for a given binding site matrix, the presence / absence of this binding site could be mapped for promoter regions in the CFCR group and 1,000 randomly sampled negative control groups, so that the statistical significance of the enrichment in a CFCR group could be measured based on a null distribution constructed with the 1,000 enrichment measurement for the negative control groups.

With this approach, we identified 65 TF binding site matrices that were enriched in at least one MixPEA identified annotation category (a CFCR gene group) (p value <= 0.01); 124 MixPEA identified annotation categories, including 95 GO biological processes and 29 pathways, were suggested with at least one enriched cis-regulatory site (Table 3). Among the binding factors of the identified cis-elements, nine TFs, including STAT, ATF3, MAZ, KROX, STAF, MRF-2, FOXO1, MYC, MAX, were under significant transcriptional regulation (FDR<0.05); the other binding factors showed high level (the transcriptional level was among the top 1/3 of all the annotated genes) but relatively steady expression, and their regulatory function are likely to be regulated post-transcriptionally (e.g. through protein modification, translocation, or protein-protein interaction).

As multiple matrices might exist for representing the binding sites of the same TF or closely related TFs, redundancy was expected among the selected matrices. To better present the independent regulatory programs, we used a logical vector to represent the presence / absence of a cis-regulatory site in each gene's promoter region, which allowed us to calculate a distance that measures how similar two cis-regulatory sites are in terms of targeting genes. Based on this distance, we built a hierarchical tree of identified cis-regulatory sites, where cis-regulatory sites located on closely related branches are having a closer relationship due to one of the two reasons: (i) high similarity between the sequences of two binding sites (e.g. binding sites of transcriptional factors in CREB/ATF family, E2F family); (ii) synergistic / competitive binding between the transcriptional factors. The later case is especially interesting, since it could infer the mechanisms, with which multiple TFs interact and alter their original regulatory function. Such interactions are usually unknown, but critical to the context-dependent transcriptional regulation. In our hierarchical tree of candidate TFs, we notice two pairs of TFs (DBP and HLF, MAZ and SP1) showed a significant overlap in their targeting genes. DBP and HLF's transcriptional regulatory function has been mainly studied in liver and kidney. It was shown before that DBP and HLF can form heterodimer, which could have enhanced the binding affinity than either of them alone. In addition to the enrichment of their binding sites, we see a steady high expression of both TFs, which supported a regulatory role of these two genes during mammary gland development, and the regulation is likely though a synergistic mechanism. The second pair, SP1 and MAZ, was also indicated by

previous study of sharing similar cis-elements. However, their interaction seem to be more context-dependent; both competitive binding and synergistic interaction was demonstrated in different cellular context. In our expression time series, both SP1 and MAZ are under moderate transcriptional regulation, but towards different direction. When the transcriptional profile of their shared putative targeting genes were analyzed, we noticed that some genes showed SP1 regulation profile, while others had MAZ regulation profile (data not shown), thus neither competitive binding model nor synergistic interaction model could be strongly supported or rejected by our data. It was previously suggested that MAZ's regulatory function is important for TATA-less promoters. It is possible that MAZ and SP1's regulatory specificity is restricted by other features in the sequence context of a promoter region, so that the same binding site sequence may not generate the same binding affinity to the two TFs.

To further inspect the regulatory roles of identified candidate TFs, we asked whether there's a preference in the biological processes being regulated. We noticed that some candidate TFs demonstrated strong regulatory specificity to certain biological process. For example, HES1 binding site was only enriched in the CFCR group related to lipid metabolism, and Tel-2 binding site was only enriched in apoptosis categories. Comparatively, many cis-elements seemed to have more general regulatory function (e.g. E2F, CREB/ATF, MAZ/SP1, and etc.). The corresponding cis-elements of these more general TFs are enriched in multiple biological processes, and are likely to contribute to the transcriptional synchronization among biological processes. When we compared

the proportions of each biological category over TF's targeting processes to the proportion over all the MixPEA identified biological processes, we still could see a strong bias in the biological processes being regulated even for these general TFs. For example, E2F binding cis-elements showed strong bias towards cell cycle processes and proliferation-related metabolism processes; SP1 and MAZ had higher enrichment in metabolism categories than the others; DBP/ HLF seemed to contribute mainly in regulating signal transduction molecules. The role of DBP/HLF in regulating signaling pathway components has not been previously addressed. We identified signaling molecules, including MAPK10, RHOU, RALGDS, PLD1 and PTK2, as the potential regulatory targets of DBP/HLF, and they all show steady high expression at late stage of acinar development (Day7-Day15), which suggested a role of the corresponding pathways in outer cell-specific differentiation. The chicken transcriptional factor, VBP, which has high protein similarity to mammalian DBP, has a pivotal role in the estrogen-dependent regulation. It is possible that DBP/HLF act as upstream response factor to female hormone and initiate transcription of signal transduction molecules critical to mammary gland development.

As an extension to the MixPEA approach, our promoter analysis is the first global investigation of transcriptional regulatory programs involved in mammary acinar in vitro development, and provided specific hypotheses of TF-biological process and TF-TF relationships. It is noteworthy that since our enrichment analysis was done by comparing an MixPEA-identified CFCR gene group to the randomly sampled negative control

groups, the TF-biological process and TF-TF relationships we suggested above are likely to be true for mammary gland development, and may not be generalizable to other tissue-type or developmental processes.

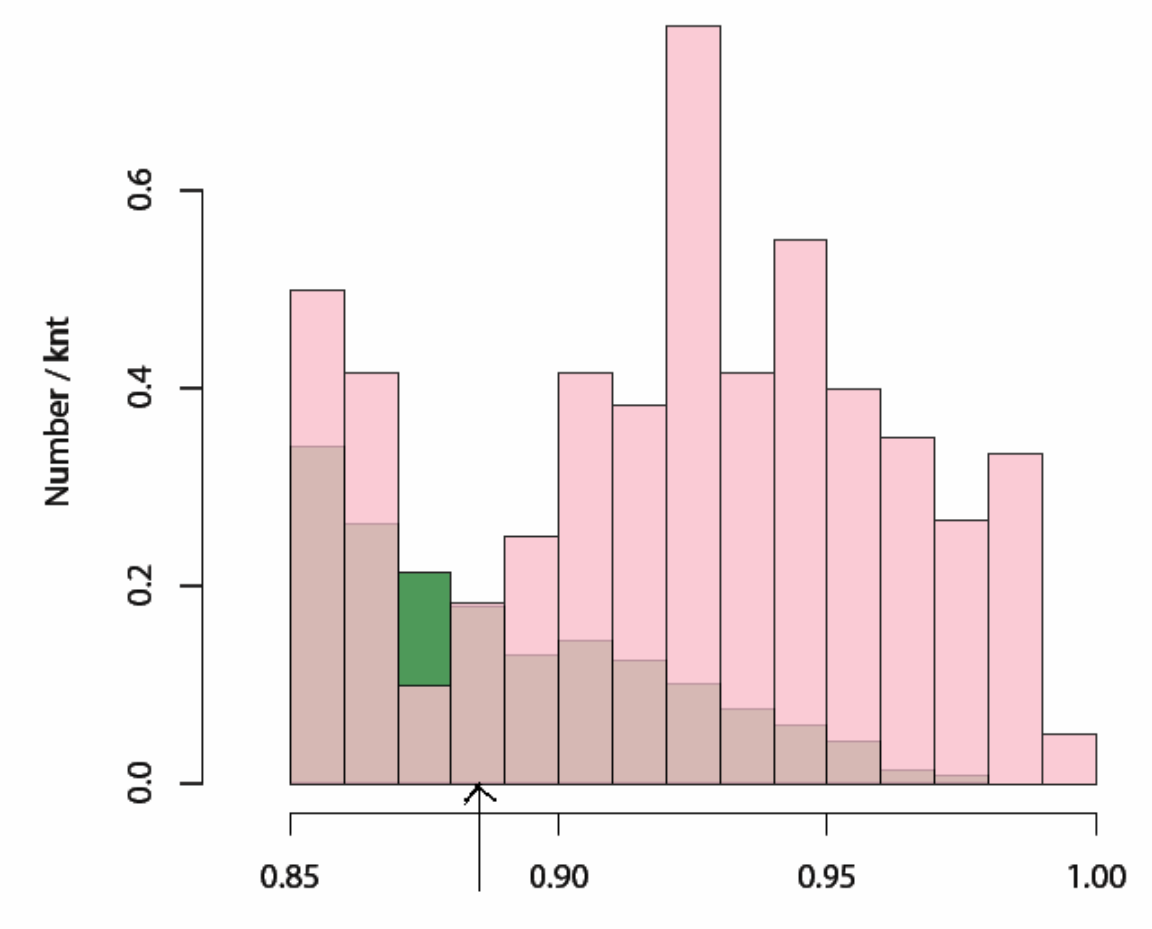


Figure 24. The distribution of TransFAC scores of Hes-1 binding sites. The pink foreground shows the score distribution over sample group and the green background distribution was from the control group. The black arrow labels the selected optimal cutoff, which achieves the smallest false negative rates without significant increase in false positive rates.

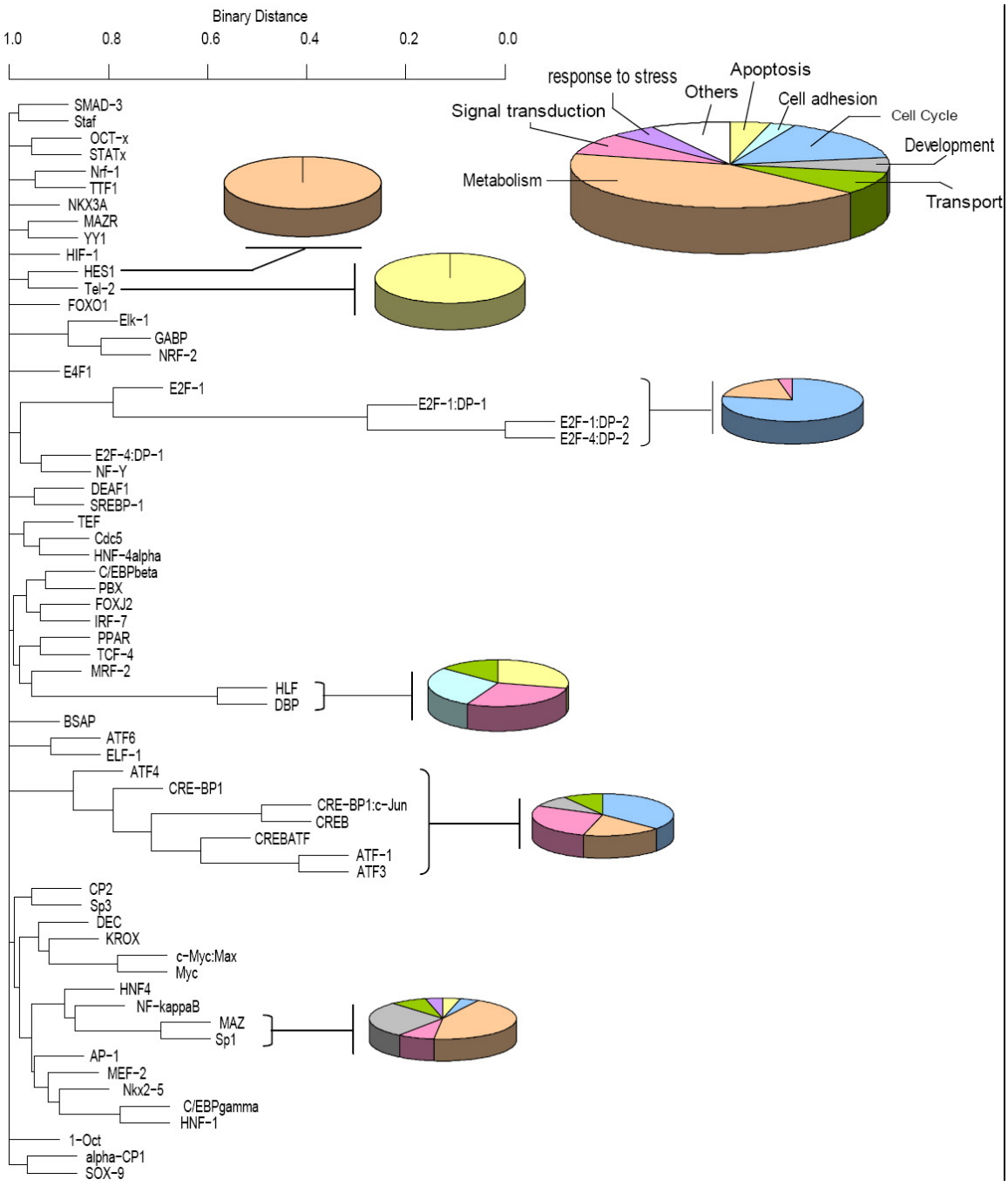


Figure 25. The overview of the transcriptional regulatory program for mammary acinar *in vitro* development.

Table 3. Candidate cis-regulatory program for MIXPEA-identified biological processes and pathways.

HLF	EnrichedCFCRGroupNumber=4	
	small_GTPase_mediated_signal_transduction|clutster_2	pvalue=0.010000
	protein_amino_acid_phosphorylation|clutster_2	pvalue=0.025000
	Apoptosis_GenMAPP|clutster_2	pvalue=0.015000
	Integrin-mediated_cell_adhesion_KEGG_GenMAPP|clutster_2	pvalue=0.015000
TEF	EnrichedCFCRGroupNumber=3	
	phosphoinositide-mediated_signaling|clutster_1	pvalue=0.020000
	protein_ubiquitination|clutster_2	pvalue=0.040000
	transcription|clutster_1	pvalue=0.045000
Nkx2-5	EnrichedCFCRGroupNumber=5	
	nuclear_mRNA_splicing,_via_spliceosome|clutster_2	pvalue=0.015000
	protein_biosynthesis|clutster_2	pvalue=0.015000
	electron_transport|clutster_1	pvalue=0.020000
	Ribosomal_Proteins_GenMAPP|clutster_1	pvalue=0.010000
	Nucleotide_Metabolism_GenMAPP|clutster_2	pvalue=0.040000
MAZR	EnrichedCFCRGroupNumber=1	
	cell_differentiation|clutster_1	pvalue=0.020000
MEF-2	EnrichedCFCRGroupNumber=7	
	signal_transduction|clutster_2	pvalue=0.020000
	sensory_perception_of_sound|clutster_1	pvalue=0.020000
	protein_complex_assembly|clutster_1	pvalue=0.035000
	Nucleotide_Metabolism_GenMAPP|clutster_1	pvalue=0.000000
	One_carbon_pool_by_folate_KEGG|clutster_1	pvalue=0.035000
	Cell_cycle_KEGG_GenMAPP|clutster_2	pvalue=0.045000
	Integrin-mediated_cell_adhesion_KEGG_GenMAPP|clutster_1	pvalue=0.045000
Sp1	EnrichedCFCRGroupNumber=20	
	development|clutster_2	pvalue=0.000000
	microtubule-based_movement|clutster_1	pvalue=0.000000
	generation_of_precursor_metabolites_and_energy|clutster_1	pvalue=0.005000
	lipid_metabolism|clutster_1	pvalue=0.010000
	cell_differentiation|clutster_1	pvalue=0.015000
	protein_polymerization|clutster_1	pvalue=0.015000
	intracellular_signaling_cascade|clutster_1	pvalue=0.020000
	transcription|clutster_1	pvalue=0.025000
	protein_transport|clutster_1	pvalue=0.025000
	immune_response|clutster_1	pvalue=0.025000
	cellular_defense_response|clutster_1	pvalue=0.025000
	glutamine_metabolism|clutster_1	pvalue=0.025000
	regulation_of_transcription,_DNA-dependent|clutster_1	pvalue=0.035000
	cell_motility|clutster_1	pvalue=0.040000
	activation_of_NF-kappaB-inducing_kinase|clutster_1	pvalue=0.040000

	nucleocytoplasmic_transport|clutster_1	pvalue=0.045000
	protein_amino_acid_ADP-ribosylation|clutster_2	pvalue=0.045000
	Cell_cycle_KEGG_GenMAPP|clutster_2	pvalue=0.035000
	G1_to_S_cell_cycle_Reactome_GenMAPP|clutster_1	pvalue=0.035000
	mRNA_processing_Reactome_GenMAPP|clutster_1	pvalue=0.040000
HNF-1	EnrichedCFCRGroupNumber=4	
	response_to_DNA_damage_stimulus|clutster_2	pvalue=0.020000
	DNA_replication|clutster_2	pvalue=0.030000
	DNA_repair|clutster_2	pvalue=0.035000
	Arginine_and_proline_metabolism_KEGG|clutster_1	pvalue=0.035000
BSAP	EnrichedCFCRGroupNumber=1	
	regulation_of_transcription,_DNA-dependent|clutster_1	pvalue=0.045000
IRF-7	EnrichedCFCRGroupNumber=13	
	transcription|clutster_1	pvalue=0.000000
	regulation_of_transcription,_DNA-dependent|clutster_1	pvalue=0.000000
	mitotic_sister_chromatid_segregation|clutster_1	pvalue=0.000000
	response_to_virus|clutster_1	pvalue=0.015000
	induction_of_apoptosis|clutster_1	pvalue=0.020000
	proteolysis|clutster_2	pvalue=0.030000
	cell_proliferation|clutster_1	pvalue=0.035000
	carbohydrate_metabolism|clutster_2	pvalue=0.035000
	nucleosome_assembly|clutster_1	pvalue=0.040000
	Glutamate_metabolism_KEGG|clutster_2	pvalue=0.000000
	Nitrogen_metabolism_KEGG|clutster_1	pvalue=0.005000
	Galactose_metabolism_KEGG|clutster_1	pvalue=0.010000
	Glycerolipid_metabolism_KEGG|clutster_2	pvalue=0.035000
PPAR	EnrichedCFCRGroupNumber=3	
	rRNA_processing|clutster_1	pvalue=0.010000
	tRNA_processing|clutster_1	pvalue=0.015000
	mitotic_sister_chromatid_segregation|clutster_1	pvalue=0.045000
SOX-9	EnrichedCFCRGroupNumber=4	
	DNA_repair|clutster_2	pvalue=0.000000
	protein_amino_acid_phosphorylation|clutster_1	pvalue=0.010000
	small_GTPase_mediated_signal_transduction|clutster_2	pvalue=0.020000
	protein_transport|clutster_1	pvalue=0.035000
ATF6	EnrichedCFCRGroupNumber=2	
	development|clutster_2	pvalue=0.020000
	immune_response|clutster_2	pvalue=0.030000
E2F-1:DP-2	EnrichedCFCRGroupNumber=13	
	cell_cycle|clutster_2	pvalue=0.000000
	regulation_of_transcription,_DNA-dependent|clutster_2	pvalue=0.015000
	mitotic_spindle_organization_and_biogenesis|clutster_1	pvalue=0.020000
	protein_ubiquitination|clutster_1	pvalue=0.025000
	protein_amino_acid_phosphorylation|clutster_2	pvalue=0.025000
	DNA_replication|clutster_2	pvalue=0.030000
	DNA_repair|clutster_2	pvalue=0.040000

	DNA_replication_initiation|clutster_1	pvalue=0.040000
	morphogenesis|clutster_1	pvalue=0.045000
	DNA_replication_Reactome_GenMAPP|clutster_1	pvalue=0.000000
	Pentose_phosphate_pathway_KEGG|clutster_2	pvalue=0.000000
	Cell_cycle_KEGG_GenMAPP|clutster_1	pvalue=0.030000
	G1_to_S_cell_cycle_Reactome_GenMAPP|clutster_2	pvalue=0.035000
E2F-1:DP-1	EnrichedCFCRGroupNumber=20	
	transcription|clutster_2	pvalue=0.000000
	regulation_of_transcription,_DNA-dependent|clutster_2	pvalue=0.000000
	cell_cycle|clutster_2	pvalue=0.000000
	regulation_of_transcription_from_RNA_polymerase_II_promoter|clutster_2	pvalue=0.000000
	DNA_replication|clutster_2	pvalue=0.000000
	DNA_replication_initiation|clutster_1	pvalue=0.000000
	nucleobase,_nucleoside,_nucleotide_and_nucleic_acid_metabolism|clutster_2	pvalue=0.005000
	protein_ubiquitination|clutster_1	pvalue=0.010000
	phospholipid_biosynthesis|clutster_1	pvalue=0.010000
	chromosome_organization_and_biogenesis_(sensu_Eukaryota)|clutster_1	pvalue=0.015000
	mitotic_spindle_organization_and_biogenesis|clutster_1	pvalue=0.015000
	DNA_repair|clutster_2	pvalue=0.020000
	regulation_of_progression_through_cell_cycle|clutster_1	pvalue=0.025000
	protein_amino_acid_phosphorylation|clutster_2	pvalue=0.025000
	nucleosome_assembly|clutster_1	pvalue=0.030000
	protein_amino_acid_phosphorylation|clutster_1	pvalue=0.035000
	Cell_cycle_KEGG_GenMAPP|clutster_1	pvalue=0.000000
	DNA_replication_Reactome_GenMAPP|clutster_1	pvalue=0.000000
	G1_to_S_cell_cycle_Reactome_GenMAPP|clutster_2	pvalue=0.000000
	Pentose_phosphate_pathway_KEGG|clutster_2	pvalue=0.000000
E2F-1	EnrichedCFCRGroupNumber=21	
	protein_folding|clutster_2	pvalue=0.000000
	transcription|clutster_2	pvalue=0.000000
	regulation_of_transcription,_DNA-dependent|clutster_2	pvalue=0.000000
	cell_cycle|clutster_2	pvalue=0.000000
	mitosis|clutster_1	pvalue=0.000000
	regulation_of_progression_through_cell_cycle|clutster_1	pvalue=0.000000
	regulation_of_transcription_from_RNA_polymerase_II_promoter|clutster_2	pvalue=0.000000
	nucleosome_assembly|clutster_1	pvalue=0.000000
	DNA_replication|clutster_2	pvalue=0.000000
	protein_amino_acid_dephosphorylation|clutster_2	pvalue=0.000000
	cell_division|clutster_1	pvalue=0.005000
	regulation_of_transcription|clutster_1	pvalue=0.005000
	DNA_replication_initiation|clutster_1	pvalue=0.005000
	chromosome_organization_and_biogenesis_(sensu_Eukaryota)|clutster_1	pvalue=0.010000

	mitotic_spindle_organization_and_biogenesis|clutster_1	pvalue=0.020000
	phospholipid_biosynthesis|clutster_1	pvalue=0.030000
	phosphoinositide-mediated_signaling|clutster_1	pvalue=0.040000
	DNA_replication_checkpoint|clutster_1	pvalue=0.040000
	Cell_cycle_KEGG_GenMAPP|clutster_1	pvalue=0.000000
	DNA_replication_Reactome_GenMAPP|clutster_1	pvalue=0.000000
	G1_to_S_cell_cycle_Reactome_GenMAPP|clutster_2	pvalue=0.000000
CP2	EnrichedCFCRGroupNumber=1	
	proteolysis|clutster_2	pvalue=0.040000
SMAD-3	EnrichedCFCRGroupNumber=1	
	positive_regulation_of_cell_proliferation|clutster_2	pvalue=0.020000
E4F1	EnrichedCFCRGroupNumber=3	
	protein_folding|clutster_2	pvalue=0.000000
	cell_cycle|clutster_2	pvalue=0.000000
	protein_amino_acid_phosphorylation|clutster_1	pvalue=0.000000
c-Myc:Max	EnrichedCFCRGroupNumber=13	
	regulation_of_progression_through_cell_cycle|clutster_1	pvalue=0.000000
	cell_division|clutster_1	pvalue=0.015000
	G2/M_transition_of_mitotic_cell_cycle|clutster_1	pvalue=0.015000
	regulation_of_transcription,_DNA-dependent|clutster_2	pvalue=0.020000
	activation_of_NF-kappaB-inducing_kinase|clutster_1	pvalue=0.030000
	cell_cycle_checkpoint|clutster_1	pvalue=0.035000
	development|clutster_2	pvalue=0.040000
	fatty_acid_biosynthesis|clutster_1	pvalue=0.045000
	G1_to_S_cell_cycle_Reactome_GenMAPP|clutster_2	pvalue=0.005000
	Fatty_Acid_Synthesis_GenMAPP|clutster_1	pvalue=0.015000
	TGF_Beta_Signaling_Pathway_GenMAPP|clutster_1	pvalue=0.015000
	Cell_cycle_KEGG_GenMAPP|clutster_1	pvalue=0.025000
	mRNA_processing_Reactome_GenMAPP|clutster_2	pvalue=0.025000
DEC	EnrichedCFCRGroupNumber=7	
	signal_transduction|clutster_2	pvalue=0.005000
	sensory_perception|clutster_1	pvalue=0.010000
	chromosome_segregation|clutster_1	pvalue=0.015000
	cell_growth|clutster_1	pvalue=0.020000
	amino_acid_metabolism|clutster_1	pvalue=0.030000
	Nuclear_Receptors_GenMAPP|clutster_2	pvalue=0.005000
	Calcium_signaling_pathway_KEGG|clutster_2	pvalue=0.025000
Nrf-1	EnrichedCFCRGroupNumber=3	
	mitotic_sister_chromatid_segregation|clutster_1	pvalue=0.000000
	protein_ubiquitination|clutster_1	pvalue=0.025000
	protein_biosynthesis|clutster_1	pvalue=0.030000
Tel-2	EnrichedCFCRGroupNumber=1	
	apoptosis|clutster_1	pvalue=0.010000
E2F-4:DP-2	EnrichedCFCRGroupNumber=13	
	mitotic_spindle_organization_and_biogenesis|clutster_1	pvalue=0.005000
	cell_cycle|clutster_2	pvalue=0.010000

	DNA_replication|clutster_2	pvalue=0.010000
	regulation_of_transcription,_DNA-dependent|clutster_2	pvalue=0.015000
	protein_amino_acid_phosphorylation|clutster_2	pvalue=0.020000
	DNA_replication_initiation|clutster_1	pvalue=0.025000
	morphogenesis|clutster_1	pvalue=0.030000
	nucleobase,_nucleoside,_nucleotide_and_nucleic_acid_metabolism|clutster_2	pvalue=0.045000
	protein_ubiquitination|clutster_1	pvalue=0.045000
	Pentose_phosphate_pathway_KEGG|clutster_2	pvalue=0.000000
	DNA_replication_Reactome_GenMAPP|clutster_1	pvalue=0.005000
	G1_to_S_cell_cycle_Reactome_GenMAPP|clutster_2	pvalue=0.020000
	Cell_cycle_KEGG_GenMAPP|clutster_1	pvalue=0.030000
E2F-4:DP-1	EnrichedCFCRGroupNumber=12	
	transcription|clutster_2	pvalue=0.000000
	regulation_of_transcription,_DNA-dependent|clutster_2	pvalue=0.000000
	nucleosome_assembly|clutster_1	pvalue=0.000000
	chromosome_organization_and_biogenesis_(sensu_Eukaryota)|clutster_1	pvalue=0.000000
	DNA_replication|clutster_2	pvalue=0.000000
	DNA_replication_initiation|clutster_1	pvalue=0.000000
	cell_cycle|clutster_2	pvalue=0.015000
	regulation_of_transcription_from_RNA_polymerase_II_promoter|clutster_2	pvalue=0.015000
	regulation_of_progression_through_cell_cycle|clutster_1	pvalue=0.020000
	Cell_cycle_KEGG_GenMAPP|clutster_1	pvalue=0.000000
	DNA_replication_Reactome_GenMAPP|clutster_1	pvalue=0.000000
	G1_to_S_cell_cycle_Reactome_GenMAPP|clutster_2	pvalue=0.000000
NKX3A	EnrichedCFCRGroupNumber=2	
	DNA_repair|clutster_2	pvalue=0.030000
	response_to_DNA_damage_stimulus|clutster_2	pvalue=0.035000
HNF-4alpha	EnrichedCFCRGroupNumber=1	
	microtubule-based_movement|clutster_1	pvalue=0.020000
HNF4	EnrichedCFCRGroupNumber=4	
	visual_perception|clutster_2	pvalue=0.020000
	immune_response|clutster_2	pvalue=0.020000
	proteolysis|clutster_1	pvalue=0.045000
	Pyruvate_metabolism_KEGG|clutster_1	pvalue=0.015000
STATx	EnrichedCFCRGroupNumber=3	
	cell_cycle_arrest|clutster_2	pvalue=0.000000
	negative_regulation_of_cell_proliferation|clutster_1	pvalue=0.000000
	cell_cycle|clutster_1	pvalue=0.010000
CREBATF	EnrichedCFCRGroupNumber=4	
	transmembrane_receptor_protein_tyrosine_kinase_signaling_pathway|clutster_2	pvalue=0.005000
	cell_cycle_arrest|clutster_2	pvalue=0.010000
	microtubule-based_movement|clutster_1	pvalue=0.020000
	mitosis|clutster_1	pvalue=0.045000

CREB		EnrichedCFCRGroupNumber=9	
		mitosis|clutster_1	pvalue=0.000000
		cell_cycle|clutster_2	pvalue=0.005000
		microtubule-based_movement|clutster_1	pvalue=0.010000
		small_GTPase_mediated_signal_transduction|clutster_2	pvalue=0.015000
		transmembrane_receptor_protein_tyrosine_kinase_signaling_pathway|clutster_2	pvalue=0.025000
		protein_transport|clutster_1	pvalue=0.045000
		protein_polymerization|clutster_1	pvalue=0.045000
		Apoptosis_GenMAPP|clutster_2	pvalue=0.010000
		Calcium_signaling_pathway_KEGG|clutster_2	pvalue=0.040000
TCF-4		EnrichedCFCRGroupNumber=2	
		positive_regulation_of_cell_proliferation|clutster_1	pvalue=0.005000
		inflammatory_response|clutster_1	pvalue=0.010000
OCT-x		EnrichedCFCRGroupNumber=2	
		nucleosome_assembly|clutster_1	pvalue=0.000000
		chromosome_organization_and_biogenesis_(sensu_Eukaryota)|clutster_1	pvalue=0.000000
	1-Oct	EnrichedCFCRGroupNumber=1	
		chromosome_organization_and_biogenesis_(sensu_Eukaryota)|clutster_1	pvalue=0.000000
ATF3		EnrichedCFCRGroupNumber=6	
		transmembrane_receptor_protein_tyrosine_kinase_signaling_pathway|clutster_2	pvalue=0.005000
		apoptosis|clutster_1	pvalue=0.010000
		morphogenesis|clutster_1	pvalue=0.015000
		cell_cycle_arrest|clutster_2	pvalue=0.020000
		mitosis|clutster_1	pvalue=0.025000
		microtubule-based_movement|clutster_1	pvalue=0.025000
MAZ		EnrichedCFCRGroupNumber=16	
		transcription|clutster_1	pvalue=0.000000
		regulation_of_transcription,_DNA-dependent|clutster_1	pvalue=0.000000
		protein_polymerization|clutster_1	pvalue=0.000000
		apoptosis|clutster_1	pvalue=0.005000
		regulation_of_transcription|clutster_2	pvalue=0.010000
		cell_surface_receptor_linked_signal_transduction|clutster_1	pvalue=0.020000
		nucleocytoplasmic_transport|clutster_1	pvalue=0.020000
		microtubule-based_movement|clutster_1	pvalue=0.020000
		cell_cycle_arrest|clutster_2	pvalue=0.025000
		fatty_acid_biosynthesis|clutster_1	pvalue=0.025000
		development|clutster_1	pvalue=0.030000
		metabolism|clutster_1	pvalue=0.035000
		lipid_biosynthesis|clutster_1	pvalue=0.035000
		Glutamate_metabolism_KEGG|clutster_2	pvalue=0.000000
		Fatty_Acid_Synthesis_GenMAPP|clutster_1	pvalue=0.015000
		Glycerolipid_metabolism_KEGG|clutster_2	pvalue=0.020000

TTF1	EnrichedCFCRGroupNumber=5	
	mitosis|clutster_1	pvalue=0.005000
	electron_transport|clutster_1	pvalue=0.010000
	fatty_acid_metabolism|clutster_1	pvalue=0.015000
	cell-cell_signaling|clutster_2	pvalue=0.030000
	lipid_metabolism|clutster_1	pvalue=0.040000
GABP	EnrichedCFCRGroupNumber=4	
	DNA_replication|clutster_2	pvalue=0.000000
	mRNA_processing|clutster_2	pvalue=0.015000
	cell_cycle|clutster_2	pvalue=0.025000
	DNA_replication_Reactome_GenMAPP|clutster_1	pvalue=0.005000
ATF-1	EnrichedCFCRGroupNumber=8	
	mitosis|clutster_1	pvalue=0.000000
	cell_cycle|clutster_2	pvalue=0.010000
	microtubule-based_movement|clutster_1	pvalue=0.010000
	cell_proliferation|clutster_1	pvalue=0.020000
	nucleosome_assembly|clutster_1	pvalue=0.025000
	chromosome_organization_and_biogenesis_(sensu_Eukaryota)|clutster_1	pvalue=0.025000
	rRNA_processing|clutster_1	pvalue=0.035000
	Apoptosis_GenMAPP|clutster_2	pvalue=0.000000
DEAF1	EnrichedCFCRGroupNumber=6	
	protein_transport|clutster_2	pvalue=0.000000
	transcription|clutster_2	pvalue=0.005000
	transcription_from_RNA_polymerase_II_promoter|clutster_1	pvalue=0.015000
	intracellular_signaling_cascade|clutster_2	pvalue=0.025000
	transport|clutster_2	pvalue=0.025000
	regulation_of_transcription,_DNA-dependent|clutster_2	pvalue=0.045000
HES1	EnrichedCFCRGroupNumber=9	
	biosynthesis|clutster_1	pvalue=0.015000
	cholesterol_biosynthesis|clutster_1	pvalue=0.020000
	steroid_biosynthesis|clutster_1	pvalue=0.025000
	regulation_of_transcription_from_RNA_polymerase_II_promoter|clutster_1	pvalue=0.030000
	induction_of_apoptosis|clutster_1	pvalue=0.035000
	sterol_biosynthesis|clutster_1	pvalue=0.035000
	lipid_biosynthesis|clutster_1	pvalue=0.040000
	Calcium_regulation_in_cardiac_cells_GenMAPP|clutster_1	pvalue=0.000000
	Cholesterol_Biosynthesis_GenMAPP|clutster_1	pvalue=0.010000
Cdc5	EnrichedCFCRGroupNumber=4	
	transcription_from_RNA_polymerase_II_promoter|clutster_1	pvalue=0.005000
	chromosome_segregation|clutster_1	pvalue=0.005000
	DNA_metabolism|clutster_1	pvalue=0.005000
	protein_amino_acid_dephosphorylation|clutster_1	pvalue=0.025000

CRE-BP1:c-Jun	EnrichedCFCRGroupNumber=9	
	mitosis|clutster_1	pvalue=0.000000
	morphogenesis|clutster_1	pvalue=0.005000
	transmembrane_receptor_protein_tyrosine_kinase_signaling_pathway|clutster_2	pvalue=0.010000
	cell_proliferation|clutster_1	pvalue=0.025000
	microtubule-based_movement|clutster_1	pvalue=0.025000
	cell_cycle_arrest|clutster_2	pvalue=0.030000
	metabolism|clutster_1	pvalue=0.045000
	cell_cycle|clutster_1	pvalue=0.045000
	Apoptosis_GenMAPP|clutster_2	pvalue=0.025000
CRE-BP1	EnrichedCFCRGroupNumber=4	
	transmembrane_receptor_protein_tyrosine_kinase_signaling_pathway|clutster_2	pvalue=0.000000
	protein_amino_acid_phosphorylation|clutster_2	pvalue=0.030000
	protein_folding|clutster_2	pvalue=0.040000
	Calcium_signaling_pathway_KEGG|clutster_2	pvalue=0.045000
FOXJ2	EnrichedCFCRGroupNumber=2	
	protein_complex_assembly|clutster_1	pvalue=0.045000
	Glycosphingolipid_metabolism_KEGG|clutster_1	pvalue=0.000000
SREBP-1	EnrichedCFCRGroupNumber=4	
	carbohydrate_metabolism|clutster_2	pvalue=0.005000
	nuclear_mRNA_splicing,_via_spliceosome|clutster_1	pvalue=0.020000
	mRNA_processing|clutster_2	pvalue=0.040000
	mRNA_processing_Reactome_GenMAPP|clutster_2	pvalue=0.020000
PBX	EnrichedCFCRGroupNumber=4	
	metabolism|clutster_1	pvalue=0.005000
	regulation_of_translation|clutster_1	pvalue=0.010000
	pyrimidine_nucleotide_biosynthesis|clutster_1	pvalue=0.025000
	Glycolysis_/_Gluconeogenesis_KEGG|clutster_2	pvalue=0.035000
KROX	EnrichedCFCRGroupNumber=8	
	cell_differentiation|clutster_1	pvalue=0.000000
	development|clutster_2	pvalue=0.000000
	immune_response|clutster_1	pvalue=0.005000
	nervous_system_development|clutster_1	pvalue=0.005000
	negative_regulation_of_apoptosis|clutster_1	pvalue=0.020000
	negative_regulation_of_progression_through_cell_cycle|clutster_1	pvalue=0.025000
	DNA_repair|clutster_1	pvalue=0.030000
	small_GTPase_mediated_signal_transduction|clutster_2	pvalue=0.045000
ELF-1	EnrichedCFCRGroupNumber=2	
	sensory_perception|clutster_2	pvalue=0.000000
	Matrix_Metalloproteinases_GenMAPP|clutster_1	pvalue=0.005000
C/EBPgamma	EnrichedCFCRGroupNumber=2	
	visual_perception|clutster_2	pvalue=0.020000

	cell_differentiation|clutster_2	pvalue=0.025000
Elk-1	EnrichedCFCRGroupNumber=4	
	nucleocytoplasmic_transport|clutster_1	pvalue=0.000000
	transcription_from_RNA_polymerase_II_promoter|clutster_2	pvalue=0.010000
	ribosome_biogenesis|clutster_1	pvalue=0.025000
	protein_biosynthesis|clutster_1	pvalue=0.045000
NF-kappaB	EnrichedCFCRGroupNumber=13	
	signal_transduction|clutster_2	pvalue=0.005000
	positive_regulation_of_I-kappaB_kinase/NF-kappaB_cascade|clutster_2	pvalue=0.005000
	G-protein_coupled_receptor_protein_signaling_pathway|clutster_1	pvalue=0.020000
	phosphorylation|clutster_1	pvalue=0.020000
	positive_regulation_of_cell_proliferation|clutster_2	pvalue=0.025000
	apoptosis|clutster_2	pvalue=0.030000
	immune_response|clutster_1	pvalue=0.030000
	cellular_defense_response|clutster_1	pvalue=0.030000
	sensory_perception|clutster_1	pvalue=0.035000
	inflammatory_response|clutster_2	pvalue=0.040000
	response_to_virus|clutster_2	pvalue=0.045000
	Translation_Factors_GenMAPP|clutster_2	pvalue=0.030000
	Biosynthesis_of_steroids_KEGG|clutster_1	pvalue=0.040000
Staf	EnrichedCFCRGroupNumber=11	
	DNA_replication_checkpoint|clutster_1	pvalue=0.000000
	proteolysis|clutster_2	pvalue=0.005000
	cell_proliferation|clutster_1	pvalue=0.015000
	mitosis|clutster_1	pvalue=0.015000
	regulation_of_progression_through_cell_cycle|clutster_1	pvalue=0.015000
	cell_cycle_arrest|clutster_1	pvalue=0.020000
	cell_differentiation|clutster_1	pvalue=0.035000
	development|clutster_2	pvalue=0.045000
	cell_division|clutster_1	pvalue=0.045000
	protein_amino_acid_phosphorylation|clutster_1	pvalue=0.045000
	Cell_cycle_KEGG_GenMAPP|clutster_1	pvalue=0.005000
Sp3	EnrichedCFCRGroupNumber=10	
	protein_folding|clutster_2	pvalue=0.000000
	positive_regulation_of_cell_proliferation|clutster_2	pvalue=0.005000
	protein_biosynthesis|clutster_2	pvalue=0.020000
	cholesterol_biosynthesis|clutster_1	pvalue=0.030000
	sterol_biosynthesis|clutster_1	pvalue=0.030000
	lipid_metabolism|clutster_2	pvalue=0.035000
	steroid_biosynthesis|clutster_1	pvalue=0.035000
	Fructose_and_mannose_metabolism_KEGG|clutster_1	pvalue=0.000000
	Calcium_signaling_pathway_KEGG|clutster_1	pvalue=0.015000
	Cholesterol_Biosynthesis_GenMAPP|clutster_1	pvalue=0.045000

MRF-2	EnrichedCFCRGroupNumber=2	
	transcription|clutster_1	pvalue=0.035000
	inflammatory_response|clutster_1	pvalue=0.040000
AP-1	EnrichedCFCRGroupNumber=4	
	regulation_of_transcription|clutster_1	pvalue=0.010000
	sensory_perception_of_sound|clutster_1	pvalue=0.010000
	proteolysis|clutster_1	pvalue=0.030000
	immune_response|clutster_1	pvalue=0.040000
YY1	EnrichedCFCRGroupNumber=2	
	DNA_repair|clutster_2	pvalue=0.000000
	response_to_DNA_damage_stimulus|clutster_2	pvalue=0.000000
FOXO1	EnrichedCFCRGroupNumber=1	
	negative_regulation_of_transcription|clutster_1	pvalue=0.000000
NF-Y	EnrichedCFCRGroupNumber=24	
	cell_cycle|clutster_2	pvalue=0.000000
	cell_division|clutster_1	pvalue=0.000000
	mitosis|clutster_1	pvalue=0.000000
	nucleosome_assembly|clutster_1	pvalue=0.000000
	chromosome_organization_and_biogenesis_(sensu_Euka ryota)|clutster_1	pvalue=0.000000
	protein_amino_acid_phosphorylation|clutster_1	pvalue=0.000000
	regulation_of_progression_through_cell_cycle|clutster_1	pvalue=0.010000
	cholesterol_biosynthesis|clutster_1	pvalue=0.010000
	steroid_biosynthesis|clutster_1	pvalue=0.010000
	sterol_biosynthesis|clutster_1	pvalue=0.010000
	lipid_biosynthesis|clutster_1	pvalue=0.015000
	isoprenoid_biosynthesis|clutster_1	pvalue=0.030000
	response_to_DNA_damage_stimulus|clutster_2	pvalue=0.045000
	biosynthesis|clutster_1	pvalue=0.045000
	Cell_cycle_KEGG_GenMAPP|clutster_1	pvalue=0.000000
	Cholesterol_Biosynthesis_GenMAPP|clutster_1	pvalue=0.000000
	Glycerolipid_metabolism_KEGG|clutster_2	pvalue=0.000000
	Biosynthesis_of_steroids_KEGG|clutster_1	pvalue=0.005000
	Galactose_metabolism_KEGG|clutster_1	pvalue=0.005000
	Terpenoid_biosynthesis_KEGG|clutster_1	pvalue=0.005000
	Glycosphingolipid_metabolism_KEGG|clutster_1	pvalue=0.015000
	Fructose_and_mannose_metabolism_KEGG|clutster_1	pvalue=0.020000
	G1_to_S_cell_cycle_Reactome_GenMAPP|clutster_2	pvalue=0.030000
	mRNA_processing_Reactome_GenMAPP|clutster_2	pvalue=0.035000
alpha-CP1	EnrichedCFCRGroupNumber=5	
	cell_cycle|clutster_2	pvalue=0.000000
	cell_division|clutster_1	pvalue=0.000000
	mitosis|clutster_1	pvalue=0.000000
	regulation_of_progression_through_cell_cycle|clutster_1	pvalue=0.005000
	Cell_cycle_KEGG_GenMAPP|clutster_1	pvalue=0.005000
VBP	EnrichedCFCRGroupNumber=4	
	electron_transport|clutster_2	pvalue=0.015000

	intracellular_signaling_cascade|clutster_1	pvalue=0.030000
	small_GTPase_mediated_signal_transduction|clutster_2	pvalue=0.035000
	protein_amino_acid_phosphorylation|clutster_2	pvalue=0.035000
NRF-2	EnrichedCFCRGroupNumber=6	
	protein_biosynthesis|clutster_1	pvalue=0.000000
	mRNA_catabolism,_nonsense-mediated_decay|clutster_1	pvalue=0.000000
	DNA_replication|clutster_2	pvalue=0.005000
	RNA_processing|clutster_1	pvalue=0.025000
	microtubule-based_movement|clutster_1	pvalue=0.040000
	DNA_replication_Reactome_GenMAPP|clutster_1	pvalue=0.005000
HIF-1	EnrichedCFCRGroupNumber=2	
	mitosis|clutster_1	pvalue=0.000000
	cell_cycle|clutster_2	pvalue=0.025000
Myc	EnrichedCFCRGroupNumber=5	
	positive_regulation_of_I-kappaB_kinase/NF-kappaB_cascade|clutster_2	pvalue=0.000000
	phospholipid_biosynthesis|clutster_1	pvalue=0.035000
	lipid_biosynthesis|clutster_1	pvalue=0.040000
	Arginine_and_proline_metabolism_KEGG|clutster_1	pvalue=0.005000
	Glycerolipid_metabolism_KEGG|clutster_2	pvalue=0.020000
C/EBPbeta	EnrichedCFCRGroupNumber=1	
	chloride_transport|clutster_1	pvalue=0.030000
ATF4	EnrichedCFCRGroupNumber=6	
	transmembrane_receptor_protein_tyrosine_kinase_signaling_pathway|clutster_2	pvalue=0.005000
	microtubule-based_movement|clutster_1	pvalue=0.005000
	protein_polymerization|clutster_1	pvalue=0.010000
	DNA_repair|clutster_2	pvalue=0.015000
	response_to_DNA_damage_stimulus|clutster_2	pvalue=0.020000
	tRNA_processing|clutster_1	pvalue=0.020000