MINING HIGH-DIMENSIONAL BIOPROCESS AND GENE EXPRESSION DATA FOR ENHANCED PROCESS PERFORMANCE

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Huong Thi Ngoc Le

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Advisor: Professor Wei-Shou Hu

July 2012

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Wei-Shou Hu, and his family for having given me tremendous support from the first day I came to the U.S. I believe that Wei-Shou and Sheau-Ping are unique in that they care for their students both professionally and personally. So much thought has been put by Wei-Shou into countless deep conversations over the years to prepare me for the best possible future. So much love has been put by Sheau-Ping into preparing all the delicious dishes we had at numerous Hu group parties and during our frequent trips to Singapore. Thank you so much for caring about me on behalf of my parents in the past six years.

I would like to thank my thesis committee members, Professor George Karypis, Professor Nik Somia, Professor Kechun Zhang, and Professor Friedrich Srienc, for taking the time to serve in my committee. Thanks, George, for having enough patience and kindness to take me step by step into the enormous field of data mining. Thanks, Nik, for teaching me so much about cloning and putting up with my endless questions. Thanks, Kechun, for sharing with me your cloning expertise so willingly. Thanks, Friedrich, for always saying hi to me in the hallway and for letting Michael help me out with so much work in the lab.

I would not have come to this point without the great lab mates that I was extremely fortunate to have: Katie Wlaschin, Anne Kantardjieff, Marlene Castro, Cornelia Bengea, Siguang Sui, Kathryn Johnson, Nandita Vishwanathan, and many others. You are my brothers and sisters who shared with me the joys and sorrows in this lonely world of graduate school. Thanks, Katie, for teaching me everything you know about cell culture in my first two years. Thanks, Anne, for helping me improve my English, build up my confidence, and communicate better. Thanks, Marlene, for sharing with me the love for classical music, nature, tacos, and so many dreams that has become the best part of my time in Minnesota. Thanks, Cornelia, for always being my big loving sister, no matter how far apart we are. Thanks, Siguang, Kathryn, and Nandita, for your sincerity and kindness that has kept me going through the long days in the lab. Special

thanks to Kathryn for taking your time to proofread this thesis and other manuscripts I wrote.

A large part of my work was enabled by the great collaboration we had with the Bioprocessing Technology Institute in Singapore, at which I have made quite a few life-long friends. Many thanks to Song Hui Chuah, Yu Xin Chin, Terk Shuen Lee, and Faraaz Yusufi for all the fun times we ate pizza and instant noodles together while working on the CHO Consortium reports. Above all, I would like to send my best wishes to Professor Miranda Yap – you are one of the strongest persons I have ever met, and I believe you will come back to us soon. We all love you.

It would be incomplete if I do not take this chance to thank all of the undergraduate students I have had over the years, who are so inspired and hard-working: Katie Scholz, Laurie Drews, Wendy Chan, Michael Srienc, and Yutin Chen. I wish you all the best in your journey to graduate school, to law school, or wherever life takes you to.

Finally, I would like to thank my big family in Vietnam for loving me unconditionally from the moment I was born. Thank you, Grandpa and Grandma, for surviving nearly a century of wars and bringing up nine children and ten grandchildren. Thanks, Mom and Dad, for taking me to this life. Thanks, my little brother, for never hesitating to share with me everything you have. Thanks, all uncles and aunts, for always encouraging me. Even though I am half of the Earth away from you, I always feel your presence in my life.

DEDICATION

This thesis is dedicated to my Grandfather,
without whom my family could not have existed.

ABSTRACT

Over the past few decades, recombinant protein therapeutics produced in cultured mammalian cells have fundamentally transformed modern medicine and improved millions of patients' lives. The drastic increase in product concentration and the number of products approved by the US Food and Drug Administration (FDA) have been attributed largely to the relentless efforts of the entire pharmaceutical community on multiple technological fronts. The remarkable advances of high-throughput genomic and process analytical tools in recent years have allowed us to extensively characterize almost all steps along a typical cell culture process. The massive amount of data generated by these technologies harbors vital information about the process, yet presents substantial challenges due to its exceptionally high dimensionality. This thesis research has applied advanced multivariate approaches to explore these sets of data and comprehend profound cellular changes during various development and manufacturing stages.

Through mining a large set of manufacturing data, we uncovered a "memory" effect, suggesting that the final outcome of a production culture is primarily affected by the early seed culture. Several parameters related to lactate metabolism and cell growth were identified as having a pivotal influence on process performance. Furthermore, transcriptome analysis of cells undergoing selection and amplification was performed using multiple statistical, clustering, and functional analysis methods. Profound transcriptional changes were discerned, upon which a combined hyper-productivity gene set involving cell cycle control, signaling, and protein processing and secretion was derived. These differentially expressed genes present promising targets for cellular modulation to enhance process performance. We further developed a novel genetic tool to engineer the expression dynamics of these genes. A large number of genes with time dynamic expression trends were identified through mining time-series transcriptome data. The promoters of these genes offer effective means to drive the expression profiles of the targets in a dynamic manner. The systems approaches outlined in this research thus hold promise to deepen our understanding of process characteristics and open new avenues for process improvement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

1 INTRODUCTION

1.1 MAMMALIAN CELL CULTURE FOR PRODUCTION OF RECOMBINANT PROTEIN THERAPEUTICS

The advent of recombinant protein therapeutics more than three decades ago has profoundly impacted modern medicine with continued discoveries and deliveries of innovative therapies for a wide range of life-threatening and serious diseases. The majority of these therapeutic proteins are produced in recombinant mammalian cells due to their unique ability to perform complex post-translational modifications, such as glycosylation, which are essential to the pharmacological activity of these biologics. Over 130 such products have gained approval by the US Food and Drug Administration (FDA) and thousands of candidate molecules are marching through different stages of the development pipeline (Leader et al. 2008). With a growth rate of 6.5% and annual sales reaching $51.3 billion in the US in 2010, the biotech sector is expected to continue to grow strongly in the foreseeable future (Aggarwal 2011). Therefore, there is an ever-increasing drive to develop genuine breakthroughs in the field of mammalian cell culture.

A typical mammalian cell culture process starts with cell line development, in which a product gene is introduced into the cells alongside other genes required for selection and amplification (Figure 1a). Single clones with excellent product concentrations (titers) and growth characteristics are selected. These clones are often subjected to amplification to increase the product gene copy number. The best clones are characterized for multiple characteristics such as growth, titers, metabolic signatures, and stability in small-scale bioreactors. During this process development step, process parameters and culture media are optimized before a few final clones with superior performance are frozen down in vials as the production cell bank. For each production run, cells from a vial are thawed and expanded in shake flasks of increasing sizes before they are transferred to seed bioreactors, also of increasing sizes (Figure 1b). At the production scale, the cells are often cultured in a fed-batch mode for approximately two weeks. The secreted protein product is harvested from the supernatant and purified using a series of centrifugation, filtering, and chromatography steps.

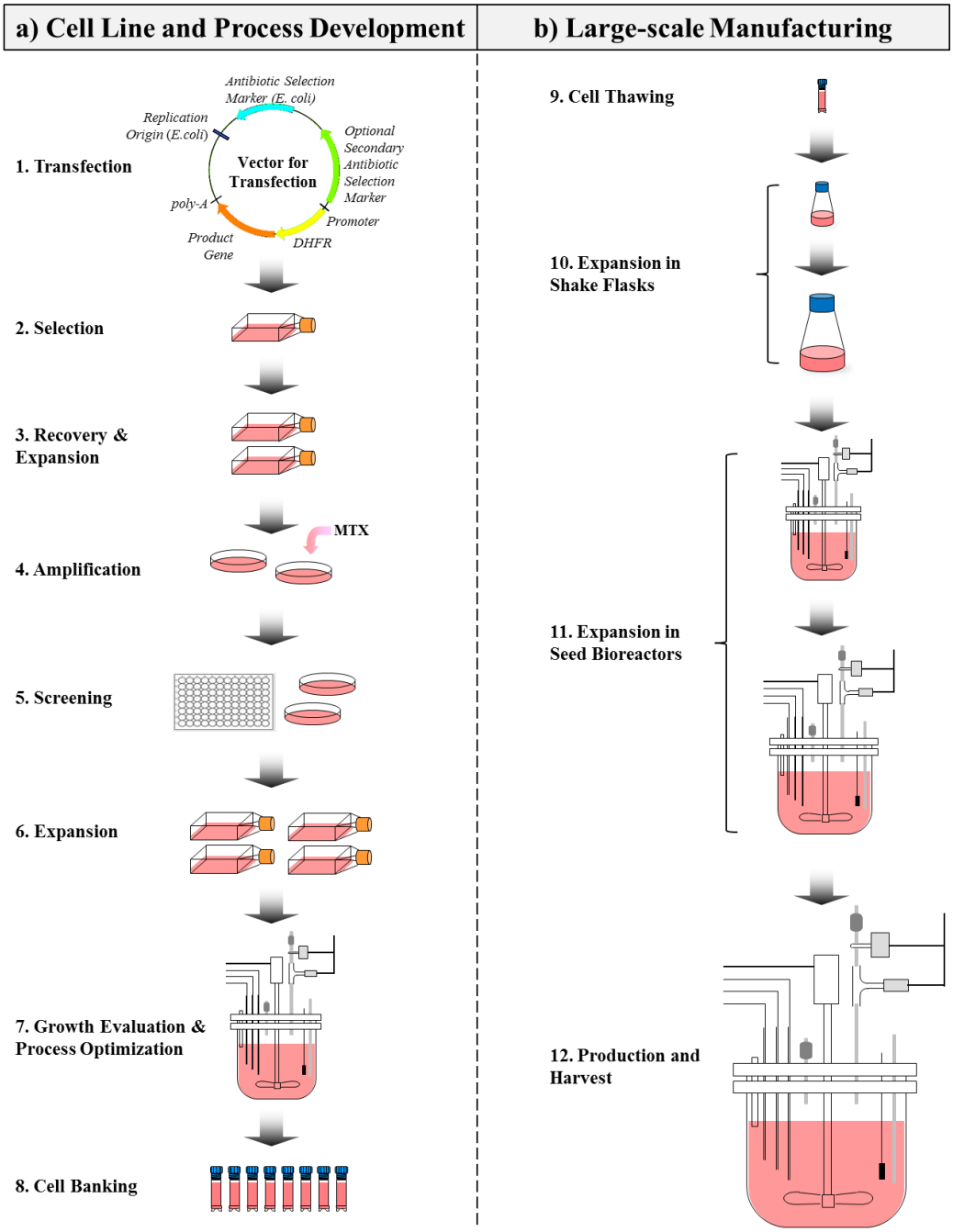


Figure 1: Overview of a typical mammalian cell culture process, adapted from (Jayapal et al. 2007).

a) Cell line and process development. Recombinant cells are generated by transfecting a vector carrying the product gene and other supporting genes. After selection and gene amplification, substantial screening is performed to identify the best clones, which are further characterized in small-scale bioreactors for growth, titer, and stability. Process parameters and culture media are optimized for the selected clones before they are banked for production.
b) After thawing and expansion in shake flasks of increasing sizes, the cells are expanded in a seed train comprising multiple bioreactors of increasing scales. At the production scale, they are often cultured under a fed-batch mode for approximately two weeks before the secreted product is harvested from the supernatant and purified.

1.2 HIGH-DIMENSIONAL BIOLOGICAL DATA

The rapid advancement of high-throughput genomic and process analytical technologies over the past few decades has enabled virtually every step along this product development and manufacturing pipeline to be extensively characterized at multiple levels. Microarray- and deep sequencing (RNA-seq)-based transcriptome approaches can be used to perform global transcriptional analysis in recombinant cells undergoing genetic alterations in cell line development and changes in culture conditions over the course of process development and large-scale manufacturing. The knowledge gained during this analysis is often extended by proteomics and metabolomics studies to investigate genome-wide changes at protein and metabolite levels, respectively. At the process level, various process analytical technologies (PAT) have facilitated real-time monitoring, control, and analysis of large-scale manufacturing processes. Each of these technologies generates a massive amount of data, which harbors valuable information about the physiological mechanisms of every step along the pipeline. Consequently, advanced data analysis approaches need to be developed and implemented to fully explore the potential that these datasets currently hold.

1.3 SCOPE OF THESIS

This thesis research focuses on the development and application of advanced multivariate data analysis approaches to high-dimensional biological data in an effort to gain mechanistic insights into mammalian cell culture processes. Two types of biological data were analyzed, namely large-scale temporal bioprocess and genome-wide transcriptome data.

A significant part of this thesis is devoted to mining a vast dataset of nearly 250 large-scale production runs acquired at a Genentech facility in Vacaville, CA, using support vector regression (SVR) and partial least square regression (PLSR). For each run, more than one million data points from nearly 150 process parameters were recorded over the course of the seed train expansion and the production run. Despite strict compliance with Good Manufacturing Practice (GMP) regarding process control and operation, there was inevitable variability in the final process outcome. Thus, through

mining such a comprehensive set of manufacturing data, we sought to evaluate the "memory" effect of the seed cultures and obtain early prediction of the final process outcome. Furthermore, we also identified critical process parameters which can potentially be used as indicators of process performance and points of intervention to render a process more robust.

In the second part of the thesis, genome-wide transcriptome analysis of cells undergoing selection and amplification was performed using significance analysis of microarray (SAM) and gene set enrichment analysis (GSEA). A custom Affymetrix microarray developed in our laboratory, with more than 61,000 probe sets representing over 15,000 unique genes, was used. The gene sets significantly enriched in the antibody producing clones during this transformation process constitute alternative paths towards hyper-productivity. Such mechanistic insights can potentially provide valuable guidelines for screening high-producers during the cell line development process.

The third part of this thesis explores time-course microarray data obtained from seventy-two samples of twelve fed-batch cultures to identify genes with time dynamic expression trends. A combination of principal component analysis (PCA) and k-means clustering was employed. The promoters of such dynamic genes provide interesting tools to control the expression of any target gene following a desired expression trend. A proof of concept has been demonstrated, and new experiments are ongoing to apply these promoters to engineering the expression dynamics of several target genes involved in regulation of central metabolism and cell growth. Such dynamic control will potentially allow cells to switch to a more efficient metabolic state at the late stage of fed-batch cultures, to maintain high viable cell density and viability for a longer period, and thus to produce greater levels of the antibody product.

1.4 THESIS ORGANIZATION

This thesis is organized into seven chapters. Chapter 2 provides a brief overview of mammalian cell culture processes used for the production of therapeutic proteins. In particular, details about cell line and process development and large-scale manufacturing are presented. In Chapter 3, background information on the characteristrics of high-

dimensional transcriptome data and a large number of multivariate approaches was provided. An example of the use of such approaches for analyzing large-scale manufacturing data is described in Chapter 4. Multivariate regression models were constructed to predict the final process outcome and to identify critical process parameters. In Chapter 5, transcriptome analysis was performed to investigate the physiological mechanisms underlying selection and amplification in a typical cell line development process. Transcriptional responses of the exogenous genes as well as other genes which potentially confer hyper-productivity traits were examined. Chapter 6 describes a novel concept of dynamic cell engineering, in which endogenous promoters with time dynamic activities can be used to drive the expression of exogenous genes following a desired dynamic trend. Such promoters were uncovered through multivariate analysis of time-series transcriptome data obtained from multiple fed-batch cultures. Finally, a brief conclusion and future directions are presented in Chapter 7.

2 BACKGROUND

2.1 SUMMARY

This chapter provides a summary of mammalian cell culture processes used for the production of recombinant protein therapeutics. More details are focused on Chinese hamster ovary (CHO) cells, the current prominent workhorse of protein production. Background information about cell line development, process development, and cell engineering is briefly described. Several aspects of these processes where opportunities for improvement exist are highlighted. Finally, the characteristics of high-dimensional bioprocess data generated from large-scale manufacturing facilities are presented.

2.2 HOST CELL LINES FOR PRODUCTION OF PROTEIN THERAPEUTICS

Despite an abundance of cultured mammalian cell lines, only a few of them currently account for the majority of therapeutic proteins produced, notably Chinese hamster ovary (CHO), baby hamster kidney (BHK), human embryonic kidney (HEK), mouse myeloma, and mouse hybridoma. The latter two were derived from *in vivo* professional antibody secretors, which are naturally equipped with fully developed secretory machinery. Thus they often achieve high productivity with merely a single copy of the product gene. In contrast, the first few, which were derived from tissues not specialized in secreting antibodies (ovary and kidney), generally required drastic gene amplification in order to produce antibody at high levels.

Among these cell lines, CHO cells currently produce nearly 70% of all recombinant protein therapeutics (Jayapal et al. 2007). Such a rise to prominence of CHO cells can be attributed to several factors. First, CHO cells have been demonstrated to perform efficient post-translational modifications, resulting in protein products compatible with human immune responses. Second, no adventitious pathogenic agents have been shown to effectively propagate in CHO cells, making them safe hosts for these products. Finally, CHO cells are easily adapted to grow in suspension, which allows volumetric scaling to large bioreactors for industrial production.

2.3 DEVELOPMENT OF CELL LINES FOR PRODUCTION OF PROTEIN THERAPEUTICS

2.3.1 Gene Delivery and Selection

Cell line development follows a generally well established process. First, a DNA molecule carrying the product gene and genes assisting selection and amplification is introduced into the host cells. In some cases, these genes can also be arranged in separate vectors which are co-transfected in unequal amounts. Typically the product gene is driven by a strong constitutive promoter, whereas the associated genes can be driven by either a similarly strong promoter or a weaker promoter to allow for more stringent selection. This DNA delivery step in protein production is often achieved using a number of non-viral mediators such as calcium phosphate, electroporation, lipofection, and polymer-mediated gene transfer. All of these methods generate a temporary disruption of the cellular membrane, allowing the DNA molecule to enter the cytoplasm. In order to generate a stable cell line, the DNA molecule has to enter the nucleus and stably integrate into the genome. Since this event occurs at extremely low frequency (less than 10^{-4}) (Mortensen and Kingston 2001), a selection step is often required to propagate stable transfectants and prevent non-transfected cells from taking over the culture.

Gene integration is largely a random event, thus not all selected cells would have the same integration site of the product gene. This variation in integration site has been known to result in profound differences in the expression level of the gene (Wilson et al. 1990). If this site resides within euchromatin, the gene is actively transcribed. In contrast, heterochromatic surroundings will significant impair the expression level. It has been shown that stably integrated transgenes can be rapidly silenced largely as a result of histone H3 and H4 hypoacetylation and loss of methylation at H3K4 (Mutskov and Felsenfeld 2004). Due to this heterogeneity of gene expression levels in transfected cells, selection is often performed in parallel with dilution cloning to isolate homologous populations, i.e., single clones. Normally, clones producing the highest antibody concentrations and growing at reasonable rates are selected.

A variety of selection systems exist, including dominant markers against antibiotics such as neomycin, blasticidin, puromycin, and hygromycin. In addition, markers which require special phenotypes are also available such as dihydrofolate reductase (DHFR) and thymidine kinase (TK), which are transferred into cells deficient in DHFR and TK activity, respectively.

2.3.2 Gene Amplification and Screening

In order to increase the expression level of the product gene, the selected cells are often subjected to gene amplification. This increase can be achieved using a number of systems, notably DHFR, glutamine synthetase (GS), adenosine deaminase, carbamylphosphate synthetase/aspartate transcarbmylase/dihydroorotase (CAD), multidrug resistance (MDR), and metallothionein. Of those, DHFR is most commonly used in CHO cells. By gradually increasing the concentration of methotrexate (MTX), which is an inhibitor of DHFR activity, the DHFR gene is amplified, resulting in co-amplification of the product gene integrated nearby. Upon MTX treatment, amplified cells frequently harbor several hundred to a few thousand copies of the product gene embedded in elongated chromosomes (Wurm et al. 1986). This step inevitably leads to heterogeneity in mRNA and protein levels of the product gene within any selected clone, since each cell responds differently to the amplification pressure. Thereby, additional dilution cloning and extensive screening are often performed to isolate the best performing clones.

In an industrial setting, hundreds to thousands of clones are typically screened for superior growth and productivity in a high-throughput manner. In addition, protein quality, metabolic behaviors, and especially stability in long-term culture are also taken into consideration. Despite containing stably integrated product gene, these cells can experience a gradual loss of product titer over time. This instability in some cases can be attributed to a decrease in the copy number of the product gene (Fann et al. 2000; Hammill et al. 2000). Decreases in transcription efficiency (Chusainow et al. 2009) and translational/secretory capacity (Barnes et al. 2004) have also been reported. Several mechanisms, including gene silencing and chromosomal rearrangement, have been suggested to possibly lead to long-term instability (Mutskov and Felsenfeld 2004;

Richards and Elgin 2002). An integration site close to the telomeres was also shown to result in stability in the absence of MTX (Yoshikawa et al. 2000).

Today, cell line development remains the most time- and resource-intensive step of an entire cell culture process. As outlined in Figure 2, this step typically takes at least 70 days with multiple empirical screenings. Thus, substantial opportunities exist for greater understanding the underlying mechanisms, which hold promise to improve efficiency and shorten the development timeline.

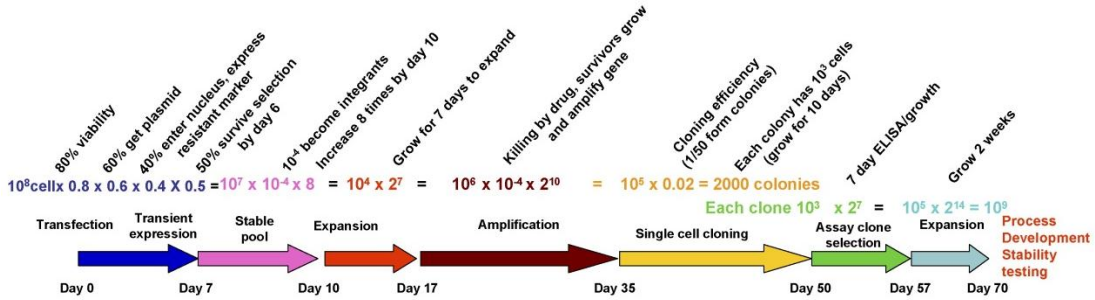


Figure 2: A typical cell line development timeline (Seth et al. 2007a).

2.4 PROCESS DEVELOPMENT FOR PRODUCTION OF PROTEIN THERAPEUTICS

2.4.1 Culture Formats

After screening, a relatively small panel of clones with superior traits is further evaluated in small-scale bioreactors. During this process development step, various parameters such as culture mode, pH, temperature, addition of supplements, especially media components, are experimented and optimized. Further genetic engineering can also be performed to improve performance of these selected clones.

Several culture modes, notably batch, fed-batch, and perfusion, are commonly used. In a batch culture, cells are growing in a fixed volume of medium until a number of essential nutrients become depleted and the viability drops. At that point, secreted product concentration reaches saturation for harvest. In a fed-batch culture, concentrated feed medium is added several days after inoculation, which prevents nutrient depletion in the culture. Thus higher cell concentration and product titer are achieved when the culture is terminated for harvest. In a perfusion culture, cells are inoculated in a batch

mode until they reach a concentration close to the desired steady state. Once concentrated medium is supplemented, an equal volume of the culture is also withdrawn to maintain the culture at a steady state. Cells in the withdrawn volume can be concentrated and recycled back to the reactor. This system allows continuous harvest of the product, yet presents a number of technical challenges that make it not as commonly used as the batch and fed-batch modes in the pharmaceutical industry. One such challenge is due to the sub-optimal condition for cell growth in continuous cultures, which does not permit the cells to enter a rapid growth period in order to produce a large amount of product. Concerns about mechanical failure and stability of the cells, especially when cell recycle is incorporated, also limit the use of continuous cultures. Thus, batch and fed-batch operating modes are still more commonly used.

Media development and optimization is arguably still the most fundamental aspect of cell culture, affecting cell growth and productivity profoundly. To date, the ranges of main components in most basal media are relatively well-defined. Depending on the cell line, several components may need to undergo optimization, which usually required statistically designed experiments (Kim et al. 1998). Serum and animal proteins are no longer widely used in cell culture media (Kallel et al. 2002). Instead, protein-free and even chemically defined media are becoming more commonplace.

2.4.2 Cell Engineering

In addition to optimization of process parameters and media components, cell engineering is often applied to further improve process performance. A number of functional pathways have been active targets of cell engineering, notably energy metabolism, amino acid metabolism, and apoptosis. The ultimate goals are to enhance peak cell concentration, lengthen culture duration, and increase product titer.

Cultured mammalian cells have been shown to uptake glucose, glutamine, and other nutrients in excess (Ronald Zielke et al. 1978). As a result, metabolic by-products such as lactate and ammonia accumulate in the medium to levels that could negatively affect cell growth and productivity. Thus, reducing lactate and ammonia accumulation has been one of the main focuses of cell engineering, and multiple efforts have been

reported. In one study, low lactate production was achieved indirectly by decreasing sugar uptake via replacing glucose with fructose and overexpressing GLUT5, a fructose transporter with lower affinity compared to the most abundant glucose transporter GLUT1 (Wlaschin and Hu 2007b). Lactate production can also be reduced by inhibiting lactate dehydrogenase A (LDH-A), an enzyme that converts pyruvate to lactate, using small interference RNA (siRNA) (Kim and Lee 2006; Le et al. 2010). Another approach was to divert the flux of pyruvate into mitochondria for entry into the TCA cycle by overexpressing pyruvate carboxylase (PCX), an enzyme that converts pyruvate to oxaloacetate (Fogolín et al. 2004; Irani et al. 1999; Kim and Lee 2007c). A combined strategy was also demonstrated by knocking down LDH-A and pyruvate dehydrogenase kinase (PDHK), which inhibits the conversion of pyruvate to acetyl-CoA (Zhou et al. 2011).

Controlling glutamine or other amino acid metabolism has been slightly more challenging as multiple amino acids share the same transporter and vice versa. To date, most of the efforts have been focused on overexpressing or de-silencing glutamine synthetase (GS) to make cells independent of an exogenous source of glutamine (Bell et al. 1995; Harris 1984).

Delaying the onset of apoptosis, or programmed cell death, has also been a target of cell engineering for several decades. Anti-apoptotic genes such as Aven, Bcl-x_L, X-linked inhibitor of apoptosis (XIAP), and E1B19K were overexpressed to improve culture performance (Figueroa et al. 2007; Figueroa et al. 2004; Sauerwald et al. 2006). Interestingly, altering apoptosis pathways was also shown to result in a metabolic switch to lactate consumption and lower production of ammonia (Dorai et al. 2009), indicating a tight link between apoptosis and metabolism.

2.5 CHARACTERISTICS OF BIOPROCESS DATA FROM LARGE-SCALE MANUFACTURING FACILITIES

Following cell line development, process optimization, scale-up, and cell banking, the production cell lines are transferred to large-scale manufacturing, at which point they are maintained in relatively small-scale bioreactors (rolling seed) to inoculate multiple

runs. Each run starts with several bioreactors of increasing scales (seed train) before the cells eventually reach the final production scale, typically about 15,000 L. These modern manufacturing facilities are highly automated in their operation and data acquisition. Hundreds of process parameters are routinely acquired and archived electronically, not only at the production scale but also throughout cell expansion in the seed train. Fluctuations in process productivity and product quality invariably occur in those production facilities. Such fluctuations or variations may exist in the same plant over time or at different plants for the same product. Understanding the root of such variations will have major economic implications for the product. Historical bioprocess data archived from these large-scale manufacturing plants thus hold much potential to provide insights for enhancing the productivity and process consistency.

Bioprocess datasets are unique in their heterogeneity. The frequency of measurement and the parameter values vary widely among different types of process parameters. Based on measuring frequency, bioprocess data can be categorized into several types, namely online, offline, raw materials, and product information. Online parameters are recorded automatically using electronic probes at extremely high frequencies of a few seconds to less than a minute for the entire production run, which typically lasts at least 20 days. Several online parameters are control parameters such as dissolved oxygen (DO), pH, and vessel temperatures, which are controlled at specific levels. For instance, DO is often set at 30% of the saturation level of oxygen in liquid medium, pH at 7.0, and vessel temperature at 37^{o}C under normal culture conditions. In addition, control action parameters are also recorded online. Examples include sparge rates of air and oxygen to control DO, and base addition and carbon dioxide sparge rate to control pH. Online data also comprise other important parameters such as vessel volume, overlay gas flow rate, and states of different valves. These valves control different ports for addition of cells, media, base, antifoam, and gas sparging.

Offline parameters, on the other hand, are recorded manually by withdrawing samples from the bioreactors. This measurement is often performed at 12 to 24 hour intervals. Most of these parameters reflect the concentrations of different nutrients and metabolites such as glucose, lactate, sodium ion, and ammonium ion. Physiological

parameters including viable cell density, viability, and packed cell volume are also measured offline.

In addition to online and offline data, information about various components of raw materials such as basal medium, feed medium, and hydrolysate are also recorded. This type of single-time point data includes lot number, the amount added from each lot, and the time stamps of each addition. Besides, documents of cell source (bank and ampoule) and cell age (time period cells are maintained in the rolling seed before they are used to inoculate a seed train) are frequently available.

The outcome of a cell culture process is often evaluated based on process quantity and quality. These values are measured at the end of the production-scale reactors, when the viability drops below a certain threshold. In addition to the secreted protein concentration in the supernatant (titer), the product quality is also a critical attribute. Commonly used criteria to assess glycoprotein quality include distribution of different glycans and charge variants.

Not only do bioprocess data vary across a wide range of measuring frequency, they also differ drastically regarding parameter value. Most parameters are quantitative with continuous values such as glucose and lactate concentrations. Several others are quantitative with discrete values such as viable cell concentration. A large number of parameters are categorical, especially lot number and valve state.

Due to this highly heterogeneous nature, bioprocess data present great challenges to the mining practice. Oftentimes, data have to be pre-processed extensively to smooth the time profiles and to estimate missing values. Selection of parameters and dimensionality reduction are also important prior to analysis. Several multivariate approaches which are commonly used for mining bioprocess data can be found in the next chapter (Multivariate Analysis of Microarray Data). A general flowchart for analyzing bioprocess data was outlined in a recent review by (Charaniya et al. 2008) and is shown in Figure 3.

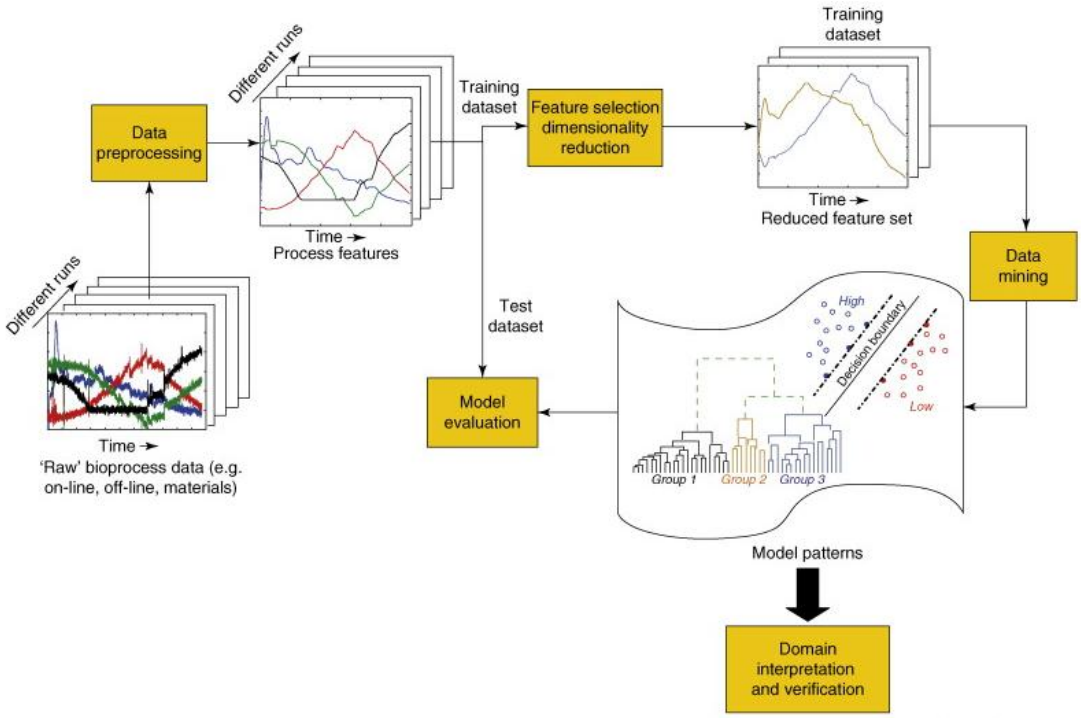


Figure 3: An approach for data-driven knowledge discovery in bioprocess databases (Charaniya et al. 2008).

3 MULTIVARIATE ANALYSIS OF MICROARRAY DATA

3.1 SUMMARY

In the past decade, DNA microarrays for transcriptome have fundamentally changed the way we study complex biological systems. By simultaneously measuring the expression levels of thousands of transcripts, the paradigm of studying organisms has shifted from focusing on local phenomena involving a few genes to surveying the whole transcriptome. DNA microarrays are used in a variety of ways, from simple comparisons between two samples to more intricate time-series studies. With the unprecedentedly large number of genes being studied, the dimensionality of the problem is inevitably high. The analysis of microarray data thus requires specific approaches. In the case of time-series microarray studies, data analysis is further complicated by the correlation among successive time points in a series.

This chapter provides a survey of the methodologies used in the analysis of static and time-series microarray data, covering data pre-processing, identification of differentially expressed genes, profile pattern recognition, pathway analysis, and network reconstruction. When available, examples of their use in mammalian cell culture are presented.

3.2 INTRODUCTION

In the past decade, genome science has drastically changed our approaches to study biosciences and has broadened our ability to harness the potential of industrial organisms for technological applications. Importantly, genome-wide gene expression profiling using DNA microarrays has become widely employed in biotechnological research. Through DNA microarrays, we are able to look at the dynamics of the transcript levels of the entire set of genes to explore the intricate relationships among the biochemical reactions, the signaling and regulation, the physiological events in the cells, and the global gene expression. In the next few years, we anticipate a greatly expanded reach of transcriptome analysis in cell culture research due to the dramatic advances in sequencing technology.

Until recently, the application of transcriptome analysis in cell culture bioprocess has been rather limited because the genome sequence information available for the most commonly used cell line, Chinese Hamster Ovary (CHO), was not extensive. With the cost of DNA sequencing drastically reduced compared to even three years ago, and the readily accessible sequencing services, one can expect that genome sequences for reference species will become available in a very near future. Furthermore, we can also expect that sequencing the genome of individual cell lines will become commonplace in a few years. Therefore, the affordability of high throughput sequencing technology will push DNA microarrays to the forefront of cell culture bioprocess characterization, along with many routinely used quantitative tools such as HPLC and ELISA. However, unlike the conventional variables typically measured in a cell cultivation process, transcriptome data are unique in its high dimensionality: each time point of measurement yields up to tens of thousands of transcript level data. In some ways, the examination of the data is like looking for patterns in a starry sky; the comparison of different datasets is as if comparing the skies in different seasons or different days.

This chapter summarizes commonly used microarray platforms and experimental design, and review methods used in differential expression analysis, profile pattern recognition, pathway analysis, and network reconstruction. In each section, an overview of the basic methodology is provided, followed by a sequence of specific modifications and associated software. Finally, several examples in which the methodology has been successfully applied are presented. When available, examples in antibody producing recombinant cell lines are emphasized.

3.3 PLATFORM OVERVIEW

Several microarray platforms are currently available, each of them offering certain advantages. As new platforms are introduced, a reduction in cost and an increase in flexibility have been observed. Microarray platforms are generally classified into two-dye or single-dye, referring to the number of fluorescently labeled samples applied to each chip.

3.3.1 Two-dye Microarrays

Two-dye microarrays were first used by Schena et al. (Schena et al. 1995) to measure the expression level of 45 *Arabidopsis* genes, and were soon followed by studies at the genome-wide level in yeast (Lashkari et al. 1997). Two-dye cDNA arrays are prepared by immobilizing long (>500 nt) cDNA probes prepared by PCR amplification onto a glass slide. cDNA microarrays allow the direct comparison of genes in two samples, each labeled with a different fluorescent dye. The raw intensities of the two dyes are indicative of the transcript levels in each sample. The probes can be designed against the genome sequence of the organism to minimize the segments which may cause cross-hybridization with transcripts from other genes. However, for mammalian cell applications, the large number of probes renders this approach very costly, as specific primers have to be designed for the amplification of specific segments of a sequence. Thus universal primers that amplify the entire cDNA region of an EST clone are more frequently used. Yet they are prone to non-specific hybridization, especially for alternatively spliced transcripts. cDNA microarrays also suffer from imprecise control of the amount of DNA immobilized on the surface, making it difficult to compare the levels of different genes in the same sample.

With the much reduced cost in oligonucleotide synthesis, cDNA microarrays are used less frequently now. In the past few years, many synthetic oligo-DNA based microarrays have evolved to be suitable for use as either single-dye or two-dye arrays. One such platform which can be used as either single-dye or two-dye is manufactured by Agilent. Similar to cDNA microarrays, short (~ 60 nt) oligonucleotides synthesized *in situ* on a glass surface. As few as individual slides are available for unique custom designs. Multiplexing – that is, the availability of testing multiple samples in a single slide – is also available in Agilent's microarrays.

3.3.2 Single-dye Microarrays

In contrast to two-dye arrays, single-dye arrays are designed to provide "absolute" measurement of the relative transcript level of each gene within a sample. With a "relative" measurement in two-dye arrays, a multiple sample comparison is cumbersome, requiring either a myriad of pairings of samples or the use of a common reference. With

an absolute measurement and one sample for each array, even meta-analysis using hundreds of microarrays can be performed.

An example of a single-dye array is that produced by Affymetrix, Inc. Affymetrix uses a photolithographic process for printing probes. Gene expression is interrogated by probe sets, which consist of eight to eleven probe pairs. Each probe pair consists of two 25-mers, one being a perfect match, the other containing a mismatch at the 13^{th} base pair. The photolithographic process, however, requires the creation of a set of masks for each array design (essentially 4 masks for each base position, thus each 25-mer will require 100 masks). The cost of generating a new set of masks limits the frequency of modifying or updating probes.

Probes on another single-dye platform, commercialized by Roche NimbleGen, Inc., are synthesized by photo-mediated chemistry using a proprietary Maskless Array Synthesizer. The use of digital mirrors creates "virtual masks", allowing for flexible designs that can be easily modified. With their ability to control the area devoted to a probe to be very small, a very large number (in the millions) of probes can be placed on a single slide. This presents an advantage for large genomes, such as those of mammalian species. For smaller genomes or with a subset of genes, array multiplexing can be implemented. Furthermore, without the use of mask, the cost of production is reduced. Making an array for only a small number of samples and frequent updating of probe design thus becomes affordable. Unfortunately, this multiplex platform will not be offered beyond 2012.

3.3.3 Other Platforms and Technologies

In addition to the glass slide chip-based microarrays, other types of arrays have been developed. One such technology is Illumina's BeadArray, which uses three micron silica beads that self-assemble in microwells with uniform spacing. In this capture technology, each bead is covered with thousands of copies of specific oligonucleotides.

With the rapid advances in DNA sequencing technologies and the decrease in sequencing cost, transcriptomes can now be analyzed by direct cDNA sequencing. In RNA-Seq, a population of RNA is converted to a cDNA library, which is then

fragmented and sequenced using high throughput technology (Wang et al. 2009b). The abundance level of a particular sequence fragment is indicative of the abundance level of the transcript from which it is derived. Unlike DNA microarrays, which can be used only to probe the expression of genes represented on the arrays, RNA-Seq detects all RNA species, including novel RNAs and alternative transcripts. It can also identify transcript boundaries, and has a much wider dynamic range, over several orders of magnitude as there is no saturation of highly expressed transcripts. However, the bioinformatics can be challenging due to extensive mapping and assembly prior to analysis.

3.4 STATIC STUDIES VS. TIME-SERIES STUDIES

Although microarrays can be used to probe transcript profiles of a large array of genes in a cell sample, most applications involve the comparison of different cell samples, either the same cell line under different conditions or different cell lines. In other words, most studies involve two or more cell samples. The use of DNA microarrays to study cell culture processes can be categorized into static or dynamic (time-series) types according to how samples are taken and compared.

Static studies compare two samples to identify differences in gene expression between them. The samples may be different cells or tissues, such as the case of comparing cell lines of different levels of antibody production (Seth et al. 2007b). In other cases, different process variables or culture conditions might be studied. The following reports provided examples wherein microarrays were used to assess the effect of cell density (Krampe et al. 2008), to study cell proliferation in protein-free medium (Spens and Häggström 2009), and to analyze the effect of hypoxic stress (Swiderek et al. 2008).

Every cell culture process constantly evolves with time, entailing various stages of culture, from the early exponential phase and exponential phase to stationary phase. In most cases, the environmental conditions change over time, either due to the culture's evolution or due to process imposed culture condition alterations such as temperature or pH shift. The gene expression profiles thus inevitably change with culture time. Static

studies offer rich information on the difference in gene expression between two conditions or two cell populations but only as a snap shot at a point of a long process.

Time-series studies sample different time points throughout the duration of the culture and aim at capturing the trends in gene expression changes originated by regulatory events and fluctuations in environmental conditions. Furthermore, the temporal information harbored in microarray time-series data also enables one to infer causality in gene regulatory networks. An aim of time-series data analysis is thus to identify genes which have different dynamic behaviors over time in the same sample or to identify the same genes whose transcripts follow different time trends under different treatments (Tai and Speed 2006). Time-series studies are particularly relevant for cellular processes exhibiting periodic behaviors, such as cell cycle and circadian rhythm, as well as other intrinsically dynamic processes such as development and differentiation. Although this type of study is abundant in yeast, roundworms, and stem cells, fewer examples have been demonstrated in antibody producing mammalian cells. In one example, gene expression time-profiles were compared between fed-batch processes yielding high and low titers using the same CHO cell line (Schaub et al. 2010). Dynamic regulation of transcription in a Human Embryonic Kidney (HEK) cell line in protein-free batch and fed-batch cultures was also unraveled (Lee et al. 2007). In addition, time-series transcriptome data was explored to elucidate cellular mechanisms leading to an increase in productivity in CHO cells under sodium butyrate treatment and temperature shift (Kantardjieff et al. 2010).

In DNA microarray studies of mammalian cell cultures, the number of differentially expressed genes and the degree of their differential expression are often lower than typical changes observed in other systems such as developmental processes or microbial cultures (Schaub et al. 2010). Using a fold-change cut-off of 1.4-2.0, and a *p*-value cut-off of 0.05-0.1, it is common to identify much less than 10% of the genes as significant. This number often decreases sharply when the fold-change cutoff is raised above 2.0 fold. For example, in studying productivity in antibody-producing cell lines, a relatively small number of genes are consistently different between high- and low-productivity clones (Seth et al. 2007b). These modest changes in gene expression thus

require careful experimental design and subsequent data analysis. This situation contrasts with most cases found in bacteria undergoing changes in nutritional or other environmental conditions, and stem cells under directed differentiation, in which often a large number of genes change their expression and many illustrate large differences in gene expression levels.

3.5 MICROARRAY EXPERIMENT DESIGN

Microarray and RNA-Seq studies can provide a wealth of information. However, even with the decrease of cost in the past few years, it is still not bargain priced. The number of conditions to be tested and the number of samples for each condition have to be planned. When single-dye microarray platforms are used, there is no limitation to which comparisons among multiple samples can be made. For two-dye arrays, however, the experimental design is crucial. With two-dye microarray platforms, one aims to measure the ratio of each gene's transcript level between two samples. When only two samples are involved, direct comparison is obtained using a single chip. With three samples, loop designs in which three arrays are used to obtain direct pair-wise comparison of the three sample pairs (1-2, 2-3, 3-1) can be applied (Kerr and Churchill 2001). An alternative, often referred to as reference design, is to hybridize each of the three samples to a common reference, and obtain indirect comparisons for each sample pair in the experiment. An often used reference is a pool of RNA, either from all samples, to ensure that the transcripts of all genes on the array are present, or from sample(s) external to the experiment. An internal reference, for instance the first sample, can also be used to directly compare some of the pairs (1-2 and 1-3) and infer comparisons for the others (in this case 2-3). The amount of available reference sample might limit the number of arrays that can be done.

Time-series microarray studies present additional experimental design challenges. Frequently, the comparison is not only among data from different time points within the same treatment but also among series under different treatments. The number of samples to be collected and their distribution in time will define the ability of the experiment to capture the gene expression dynamics. Sample collection frequency should be high enough to capture the dynamics of genes with periodic behaviors or propensity for

sudden changes in expression. This approach, however, might result in a very large number of samples, which is not always feasible due to the cost or the amount of work involved (Wang et al. 2008). If critical changes are suspected between the originally analyzed time points, additional microarrays can be performed. This is possible if samples were collected at intermediate time points. Another possibility is to fill these gaps using quantitative PCR measurements of transcripts of the target genes.

The general steps of analyzing gene expression data from microarrays is shown as a flowchart in Figure 4. First, raw data are filtered to eliminate absent probes using intensity and/or detection *p*-value cutoffs. Filtered data are further normalized to generate a baseline for comparison across samples. Time alignment, log transformation, and scaling can be performed if necessary. Once the data have been properly processed, genes exhibiting differential expression can be identified using multiple statistical approaches. These significant genes are often further analyzed in a pathway/network context or using clustering tools to infer the biological meanings of differential expression.

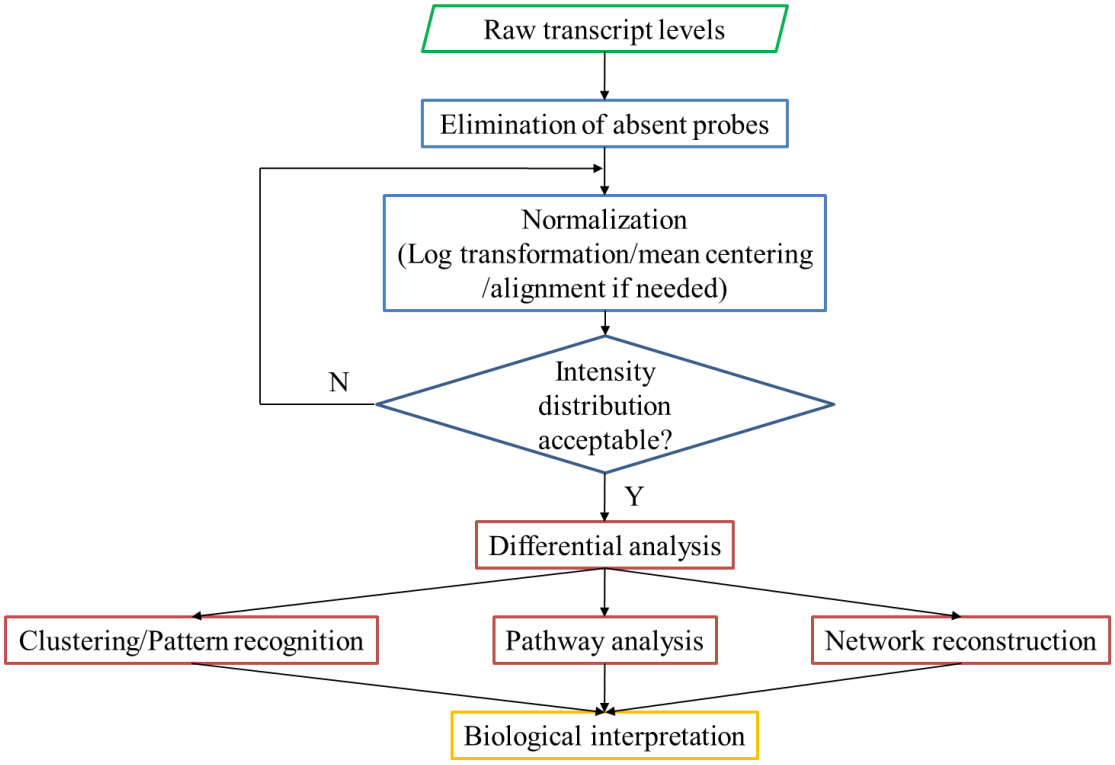


Figure 4: General analysis flowchart for microarray data.

Raw expression data are often filtered to eliminate absent probes. The filtered data are subsequently normalized and processed using different methods if necessary. Genes exhibiting differential expression are identified using statistical tools. In addition to clustering, further analysis can be performed in a pathway/network context to finally interpret the biological meaning of differential expression.

3.6 DATA PRE-PROCESSING

3.6.1 Normalization, Transformation, and Scaling

Gene expression levels measured using DNA microarrays are subject to a number of systematic biases, and hence should be globally adjusted (or normalized) to attain a common basis for all the microarrays to be compared. These variations in gene expression measures are often the result of differences in starting amounts of RNA, labeling, hybridization, and scanning efficiency (Quackenbush 2002). Normalization is thus a necessary step regardless of the platform used or whether the experiment involves static or time-series samples. Different normalization methods (based on different sets of assumptions) often give different quantification. Most normalization methods assume that the microarray contains a large and unbiased set of genes. Furthermore, the number

of differentially expressed genes is considered to be relatively small compared to the total number of genes present on the array. As a result, this differential expression does not affect the overall distribution of gene expression levels in each sample.

Linear and quantile normalization are the most commonly used normalization methods in microarray data processing. Linear normalization is often applied when gene expression measures in all arrays have similar distributions but different median values. Given the assumption that equal amounts of RNA are used in each sample, a normalization factor is calculated as the ratio of the median gene expression levels in two samples (Quackenbush 2002). All gene expression measures are subsequently scaled using this factor such that these two samples have the same median gene expression level after normalization. A target median value can also be defined to linearly scale multiple samples. Thus linear normalization is conceptually simple, yet applicable to most cases wherein the assumptions stated above are satisfied. However, possible lack of linearity between fluorescence intensity and the amount of DNA or RNA hybridized could introduce errors when linear normalization is applied.

Quantile normalization, on the other hand, assumes that all samples have the same gene expression level distribution (Bolstad et al. 2003). Gene expression measures are adjusted such that each sample follows the same distribution, which is assumed to be the average distribution of all samples. This normalization method is frequently used to correct the gene expression level distribution in single-dye and two-dye arrays when genomic DNA is used in one channel. Sometimes, a drastic change in cell physiology may occur, causing a major shift in gene expression profiles. In such cases, the use of quantile normalization might not be appropriate. For example, as stem cells differentiate or cells enter different phases of growth, their transcriptional responses or cellular RNA composition may change drastically. Large variations in cellular RNA composition among samples violate the assumption that all samples have the same gene expression level distribution.

It is important to note that, in most experimental protocols, the amount of total RNA (in the case of prokaryotic samples) or poly(A)-tailed transcripts (in the case of eukaryotic samples) applied to each array is kept equal, thus normalization methods only

adjust the data to equal quantities of RNA. However, the RNA content per cell does not always remains constant under different conditions. Fast growing cells have far more RNA than cells in stationary phase, thus total RNA content per cell varies. It is thereby important to know whether differential expression calls are based on per cell or per unit amount of RNA.

After normalization, the data are usually log-transformed. The variance, which is inherently large in microarray data, is reduced in log-transformed data. Normalized gene expression values can also be scaled to a mean or median value of zero. This is equivalent to centering the gene expression level distribution over zero (mean- or median-centering). Additionally, a standard deviation of one can be achieved using *z*-transformation. These data pre-processing steps can be performed using several software including Expressionist, GeneSpring, and R packages such as *affy*, *limma*, *beadarray* and *oligo*. Although data normalization, transformation and scaling have become routine, these steps remain vital to all subsequent stages along the analysis pipeline for gene expression data.

3.6.2 Time Alignment

When comparing time-series experiments, it is important to ensure that the starting cell populations in different treatments are identical, or at least as similar as possible. Under some conditions, variability is difficult to eliminate, resulting in somewhat different kinetic profiles even among biological replicate cultures. When applying microarray to time-series studies, the aim is to identify genes whose transcript dynamics change beyond the fluctuations in biological replicate cultures, especially changes which can be attributed to experimental treatment. In assessing the similarity or difference between two cultures under different treatments, a direct comparison of time profiles is an obvious first approach. This is sound in the cases that the trends of growth and other growth-related variables (such as chemical profiles) are mostly identical. Often growth and other culture indicators reveal difference, strongly hinting that the identical time points in two cultures may not correspond to identical “culture stages”. In other words, the time frame of one culture has shifted from the reference time frame of the other culture. Direct comparison of time profiles of gene expression may give rise to

many falsely identified genes with different kinetic behaviors. Thus these potential misalignments should be identified and corrected prior to analysis.

The change in time dynamics could be global, i.e., all the transcripts change their temporal profiles similarly. This change may also be segmented and local, i.e., only some sets of genes change coordinately apart from the rest of genes or different sets of genes change their dynamics differently. Such asynchronous behaviors need to be dealt with using some forms of time alignment. Asynchronization between transcriptome time profiles appears in multiple forms, which can be largely divided into four types: frame shift, elastic compression or expansion, and time flip (Mehra et al. 2006).

Frame shift occurs when one of the series experiences a lag phase with respect to the others. If the growth rate differs significantly among the series, their gene expression profiles may display elastic compression or expansion. Examples of frame shift and expansion are shown in Figure 5a. These types of asynchronization are often adjusted globally. In addition, changes in a few subsets of genes can result in a flip in time order between different subsets of genes (Figure 5b). This time flip suggests the existence of multiple biological clocks controlling varied cellular processes in the experimental system and thus requires local alignment. As a result, when multiple treatments are being compared, gene expression data sets should be examined and, if necessary, properly aligned before subsequent analyses can be performed.

Conceptually, aligning time-series microarray data entails matching two patterns by locally compressing, expanding, or translating one with respect to the other such that their similar characteristics are aligned without altering the ordering of each sample. This can be performed on either the continuous representation of each series or the discrete values of gene expression. The alignment between time series can be achieved at a global level or at a local level to allow different subsets of genes to follow varied biological clocks.

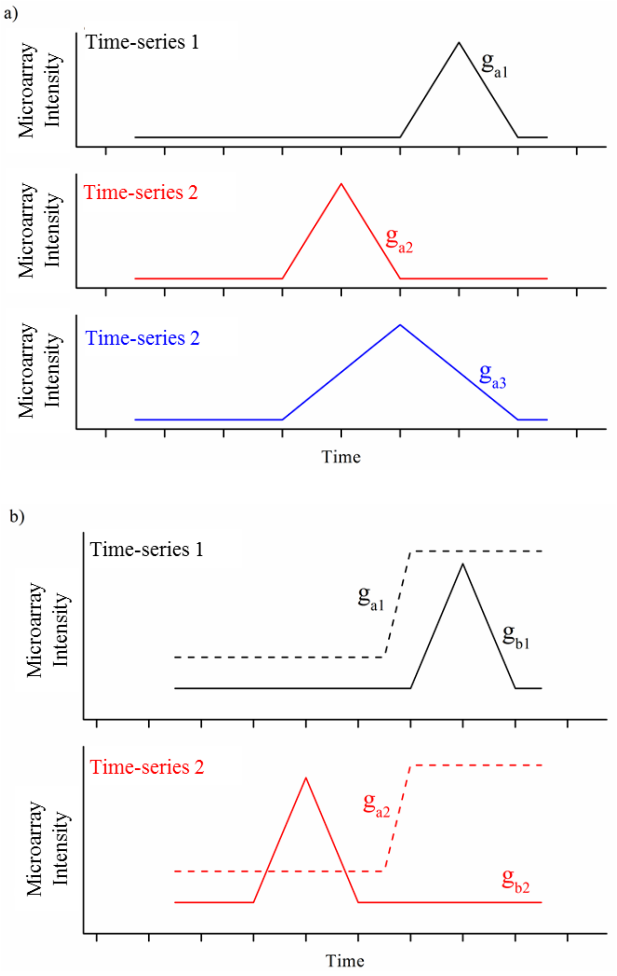


Figure 5: Possible forms of time asynchronization in different series.

a) Expression profile of gene g_a in three different series. The expression profile in series 2, g_{a2}, shows a shift with respect to series 1, g_{a1}. The expression profile in series 3, g_{a3}, shows an expansion with respect to g_{a1}. These types of asynchronization can be adjusted globally.

b) Expression profiles of two genes, g_a and g_b, in two different series. Gene g_a displays the same expression profile in the two series. The peak observed in the expression profile of gene g_b in series 1 appears at an earlier time point in series 2. This time flip often requires local adjustment.

An example of global alignment is the B-spline based alignment method, which represents each gene expression profile as a spline curve of multiple low-degree polynomials (Bar-Joseph et al. 2003). To align different time series, one of the series is chosen as the reference, and the time points of the other series are mapped to the reference series by stretching and shifting the continuous representation of gene profiles. This method is particularly suited for long time series (e.g., $\geq$ 10 time points) (Bar-Joseph 2004). The use of B-splines for alignment was demonstrated by aligning three yeast time

series that begin in different phases and occur in different time scales (Bar-Joseph et al. 2003).

A second example for global alignment, dynamic time warping (DTW), involves non-linear mapping between discrete time points of two series along the time dimension such that the distance between them is minimized (Sakoe and Chiba 1978). In the case of transcriptome time series, the overall distance between the two series is computed as the weighted sum of distances contributed by all genes. The use of a weighting factor for each gene allows higher contribution to the overall distance measure to be given to genes with consistent expression profiles across two treatments, or to genes important to the biological activities being considered. In Figure 6, the algorithm is exemplified with two genes (g_a and g_b) in two different series. The weighting factors (w_a and w_b) indicate the contribution of each gene to the final adjustment. g_a with a less dynamic profile thus has a lower weighting factor ($w_a < w_b$). In addition to alignment of transcriptome data, DTW has also been used to synchronize offline and online data of batch processes (Kassidas et al. 1998; Ramaker et al. 2003).

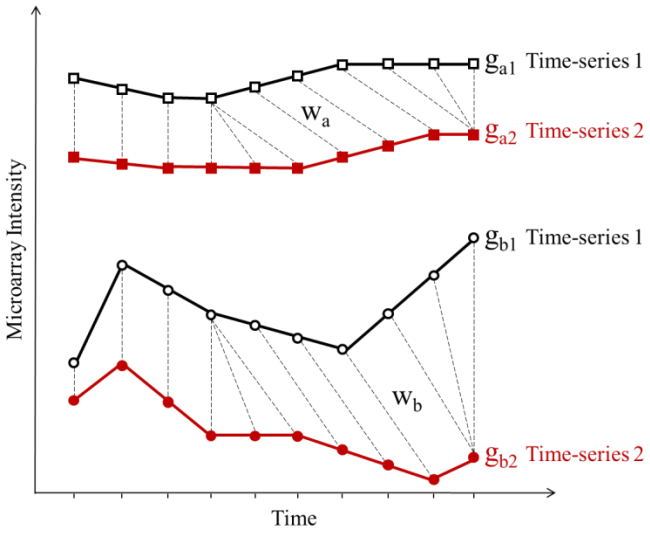


Figure 6: Alignment of gene expression profiles using DTW.

Two genes (g_a and g_b) are shown in two different series. For each gene, discrete time points in series 1 are mapped to those in series 2. The weighting factors (w_a and w_b) indicate the contribution of each gene to the final global adjustment. Gene g_a with a relatively flat profile is given a lower weight ($w_a < w_b$). The same alignment is imposed on both genes.

Global alignment algorithms assume that all genes share the same alignment, that is, that all genes were affected in the same manner. The existence of multiple biological clocks within the same cell, however, can result in sets of genes affected independently. In other words, genes in one set correspond to genes that follow a particular biological clock, sharing the same alignment, but they need to be warped separately from the rest of the genes. Recently, Smyth et al. (Smith et al. 2009) have proposed an algorithm capable of identifying sets of genes that present similar alignments when aligned independently. The resulting sets include genes that follow similar warpings, even though their expression profiles might be very different.

3.7 IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

After transcriptome data have been pre-processed, a number of statistical approaches can be used to identify differentially expressed genes. Commonly used analytical methods for static transcriptome data include *t*-tests, ANalysis Of VAriance (ANOVA), Significance Analysis of Microarray data (SAM), and LInear Models for MicroArray data (*limma*). These methods are not all directly applicable for dynamic studies involving a chronological set of samples collected over time, since a change in the time order will result in a different statistical inference (Storey et al. 2005). Recently several methods based on regression, ANOVA, and Bayesian models have been adapted to handle time-series microarray data. In addition, a distance calculation approach has also been proposed for identification of kinetically differentially expressed genes.

3.7.1 Statistical Analysis of Gene Expression Data

The estimates of gene expression levels provided by microarray data are generally prone to two types of errors – systematic and random errors. Systematic error resulting from several factors such as RNA concentration measurement or dye-labeling efficiency can give rise to a systematic bias in the expression level estimates of all genes on the same array. This bias is often corrected using one of the normalization methods presented in the previous section of data pre-processing. Random error in the measurement of gene expression levels arises from random fluctuations in other steps, for

instance array scanning. Inferential statistics is used to ensure the observed change in gene expression did not occur by random chance.

Inferential statistics is applied to microarray data by invoking a null hypothesis. The null hypothesis holds true when all samples have the same average expression value for the gene of interest. Conversely, if the gene is expressed at a different level in at least one sample, the alternative hypothesis becomes valid. In order to assess the validity of either hypothesis, a test statistic is often estimated as the ratio between the change in a gene's expression values among samples and the variability in those measurements. Further, a *p*-value computed using this test statistic is compared to an acceptable significance level α. The smaller the *p*-value is compared to α, the stronger the evidence is against the null hypothesis, and in support of the gene being differentially expressed in at least one sample.

In a typical microarray experiment, tens of thousands of genes are tested simultaneously, and a large number of them are likely to be identified as differentially expressed. Even with a small *p*-value such as 0.01, which is normally considered to be rather stringent, a significant number of those identified as differentially expressed might be called so by random chance. For example, if 1,000 genes out of 10,000 in total are identified as differentially expressed, each with a *p*-value < 0.05, then 500 of these 1,000 genes might be identified by chance. One way to control the potentially high error rate is to set each gene's *p*-value to an *n*-fold lower significance level, α/n, where n is the total number of genes. This method is often referred to as the Bonferroni correction for the family-wise error rate – (FWER) (Bonferroni 1936). However, this correction imposes an extremely stringent criterion. In the previous example, the *p*-value will have to be set at less than 0.0000005. This strict criterion would likely result in failure to identify the majority of genes that are indeed differentially expressed. An alternative is to control the number of false positives among the number of genes declared as differentially expressed rather than the total number of genes. This statistic, referred to as false discovery rate (FDR), is less stringent than the FWER and thus offers more power than the FWER to detect differential expression (Benjamini and Hochberg 1995). Therefore, in multiple hypothesis testing, FDR is often used in place of *p*-value.

3.7.1.1 Statistical Analysis of Static Gene Expression Data

A variety of methods are available for hypothesis testing. A *t*-test is often used when only two samples are compared to test for differential gene expression. When three or more samples are involved, ANalysis Of VAriance (ANOVA) is recommended to avoid performing multiple *t*-tests, which will most likely result in an increased false positive rate. Both methods assume the expression levels of a gene in different samples follow a normal distribution. When this assumption does not hold true, non-parametric tests including the Wilcoxon rank-sum test and a permutation-based test are often the methods of choice.

3.7.1.1.1 *t*-test

t-tests are considered the simplest statistical methods to identify differentially expressed genes. A *t*-statistic is calculated as the ratio between the difference in gene expression levels of two samples and the pooled variance. Furthermore, a degree of freedom is calculated from the sample sizes – with more penalties if the two samples have unequal variances (Welch's *t*-test), and no penalties if the assumption of equal variances holds true (Student's *t*-test). A *p*-value, which can be obtained using the *t*-statistic and the degree of freedom, is compared to a pre-defined significance level α to detect differential expression. *t*-tests can be easily performed in Microsoft Excel, several R packages, and a variety of software including Spotfire and Expressionist.

Gene expression responses during metabolic shift in a hybridoma cell culture have been investigated using the Student's *t*-test on cDNA microarray data (Korke et al. 2004). More than 120 probes were identified as changing their expression levels (fold-change ≥ 1.4 and *p*-value ≤ 0.1) when the cells were shifted to the lactate consumption state. Another example involves the survey of global gene expression changes in a recombinant antibody producing Chinese Hamster Ovary (CHO) cell line and a mouse hybridoma cell line under sodium butyrate treatment (De Leon Gatti et al. 2007). Using a fold-change cutoff of 1.4 and a *p*-value cutoff of 0.05, most transcripts were found to be expressed at similar levels in both cell lines, indicating the transcriptional responses under exposure to sodium butyrate are rather conserved.

3.7.1.1.2 Analysis of Variance (ANOVA)

When more than two samples are involved, single-factor ANOVA is often used. The overall variance in gene expression among different samples is partitioned into separate sources of variations. The total variation, as evaluated by sum of squares (SS_{Total}), arises from two sources – the actual differential expression among these samples ($SS_{Treatment}$) and the random error (SS_{Error}). The mean sum of square for treatment ($MS_{treatment}$) and that for error (MS_{error}) can be estimated by dividing each SS by the corresponding degree of freedom. The quotient of these two MSs is taken as the *F*-statistic, which further provides a *p*-value for inference of differential expression.

When the experiment involves several variables, or factors as they are known in ANOVA, and one wishes to segregate the effects of those factors, multiple-factor ANOVA is used. Based on the same working principles described above, multiple-factor ANOVA also partitions the total variation into different sources – the actual effect of each experimental factor, their interactions, and the random error. A *p*-value for each term can be derived similarly, and whether these factors significantly affect the change in gene expression can thus be concluded. Both single-factor and multiple-factor ANOVA can be performed easily using Microsoft Excel, as well as several R packages.

Variation in gene expression within and between two populations of the genus *Fundulus* was uncovered using ANOVA on $\log_2$ normalized microarray data of 907 genes (Oleksiak et al. 2002). More than 160 genes were differentially expressed among individuals within a population, whereas only 15 genes differ between populations, suggesting that substantial natural variation exists in gene expression. A linear ANOVA model was also fitted to the expression levels of more than 3,000 genes expressed during embryonic development of six *Drosophila* species (Kalinka et al. 2010). More than 80% of genes best fit to models incorporating stabilizing selection, and maximal similarity is observed during mid-embryogenesis rather than early or late stages of development. This result thus supports the developmental hourglass model, and the hypothesis that natural selection acts to conserve gene expression patterns during the phylotypic period.

3.7.1.1.3 Significance Analysis of Microarray (SAM)

Similar to *t*-tests, SAM also calculates a "relative difference" (d), which resembles the ratio between difference in average gene expression values and the pooled variance in two treatments for each gene (Tusher et al. 2001). The expression levels in all replicated samples of these two treatments are then permuted, and an average "relative difference" over these permutations (d_E) is estimated. For the majority of genes, which are assumed not to be differentially expressed, the average difference obtained from permutation (d_E) is largely the same as the observed one (d). If the discrepancy between d_E and d exceeds a threshold, the gene is considered differentially expressed. In order to calculate the FDR for each gene, two horizontal cutoffs are defined – one as the smallest observed difference of up-regulated genes, and the other as the least negative of down-regulated genes. The average number of genes with d_E exceeding these cutoffs in all permutations can be considered as the number of false positives, and is used to assess FDR. A convenient Microsoft Excel-add-in for SAM is available, and the packages *siggenes* and *samr* in R are also publicly accessible.

The advantage of SAM over other statistical methods was demonstrated when examining the transcriptional responses of the human lymphoblastoid cells under irradiation (Tusher et al. 2001). Thirty-four genes were identified as significant at an FDR of 12% using SAM, compared to more than 60% using other methods. In another example, SAM was used to identify about 400 genes contributing to the impaired differentiation capacity of murine neural stem cells (NSCs) defective in p53 and PTEN genes (Zheng et al. 2008). The majority of genes involved in cell cycle regulation were also found to be significantly down-regulated when HeLa cells were transfected with siRNA against PHF8, an H4K20me1 demethylase (Liu et al. 2010a).

3.7.1.1.4 Linear Models of Microarray Data (limma)

In this approach, a linear hierarchical model with arbitrary coefficients and contrasts across multiple samples for each gene is developed (Lonnstedt and Britton 2005; Smyth 2004). Further, marginal distributions of the observed statistics are used to estimate the hyperparameters under consistent and closed forms. In addition, the ordinary *t*-statistic can be replaced by a moderated one, which implicitly results in

shrinkage of all gene-wise variances into a common value. This moderate *t*-statistic follows a *t*-distribution with augmented degrees of freedom, and thus can be extended for multiple-sample comparisons by using the corresponding *F*-statistics. The R package *limma* is publicly available.

Transcriptional responses upon restoration of p53 in adenocarcinomas were revealed using *limma* (Feldser et al. 2010). p53-restored samples were shown to cluster with adenomas rather than carcinomas, suggesting that adenocarcinoma cells can be specifically removed from the tumors. *limma* was also used to compare gene expression signatures between cultured thymic epithelial cells (TECs) and multi-potent hair follicle (HF) stem cells (Bonfanti et al. 2010). About 120 genes were identified to be differentially expressed between these two samples with a fold-change cut-off greater than four and a *p*-value less than 0.001.

3.7.1.2 Statistical Analysis of Dynamic Gene Expression Data

Time-series transcriptome data offer a great advantage to explore transcription as a dynamic process, yet their analysis is more complicated than analyzing multiple samples unrelated in time. Transcriptional responses at a certain time point often carry information about cellular behaviors in previous stages. Thus samples within a series are mutually dependent, and should not be analyzed using traditional statistical approaches. Rather, methods taking this interdependency into consideration, such as Extraction of Differential Gene Expression (EDGE), Microarray Significant Profiles (maSigPro), ANalysis Of VAriance – Simultaneous Component Analysis (ANOVA-SCA), and multivariate Bayesian models, are more suitable. The number of time points in each series, the number of series, and the availability of replicates will guide the selection of an algorithm to use in data analysis. This analysis can become even more challenging if the sampling frequency is not uniform across multiple series.

3.7.1.2.1 Extraction of Differential Gene Expression (EDGE)

In EDGE, differential analysis is also approached as a hypothesis testing problem. The null hypothesis is that a gene's expression does not change either over time within a single treatment or across multiple treatments (Storey et al. 2007; Storey et al. 2005).

The expression profile of each gene is modeled using a p-dimensional basis, usually a p^{th}-order polynomial, or a natural cubic spline function. The parameters of these functions are then estimated by minimizing the sum of squared errors (SSE) between the model-fitted expression values and the corresponding actual ones. The parameterization of gene expression profiles allows the hypothesis testing to be performed by comparison of the fitted parameters. As such, an F-statistic is calculated for each gene to reflect the relative difference in SSE of the model-fitted gene expression profiles under the null and the alternative hypotheses, respectively. This statistic is used alongside a null distribution generated using a resampling method to estimate a q-value, which accounts for the FDR incurred in multiple hypothesis testing (Storey and Tibshirani 2003).

The open-source software EDGE (Leek et al. 2006) has facilitated the use of this methodology in analyzing time-course gene expression data. Differential expression can be surveyed along the time axis within each treatment or across multiple treatments. EDGE was used to define the transcriptomic signatures of aging in several tissues in *Drosophila melanogaster* (Zhan et al. 2007). In a mouse model, a complex transcriptional hierarchy comprising more than one thousand genes regulated during endocrine differentiation was also identified using EDGE (White et al. 2008).

3.7.1.2.2 Microarray Significant Profiles (maSigPro)

Microarray Significant Profiles, maSigPro (Conesa et al. 2006), uses a two-step regression approach to identify differentially expressed genes in time-series microarray data. Single or multiple time series can be analyzed, with multiple time series being analyzed directly instead of performing multiple pair-wise analyses. This methodology not only detects kinetically differentially expressed genes, but also uncovers changes in gene expression trends. In the first step of gene selection, expression data are fitted using a global regression model which considers all experimental variables and their interactions. If there are n groups, $(n - 1)$ dummy variables are defined. Each dummy variable allows the distinction between each group and the reference group. Further, an ANOVA table is generated for each gene. If the gene shows differences between any group and the reference group, the regression coefficients will be statistically significant as determined by an F-statistic and its associated p-value. In the second step of variable

selection, the best model for each gene is obtained using a stepwise regression approach. The variables that best fit the data represent the time effects and their interactions with the dummy variables. For finding those genes with significant differences in group *x* with respect to the reference series, the genes with significant coefficients for the dummy variable ($x - 1$) are selected.

The package maSigPro is available in R and includes several tools for result visualization. In addition, it is part of the oneChannelGUI package (Sanges et al. 2007), which provides a graphical interface for the analysis of Affymetrix microarrays and was included in the popular software Gene Expression Pattern Analysis Suite (GEPAS) (Tarraga et al. 2008). An extension of maSigPro, maSigFun (Nueda et al. 2009), is used to fit regression models for genes with the same functional class and for the functional assessment of time-course microarray data. maSigPro has also been implemented in Corra (Brusniak et al. 2008), an R package devoted to the analysis of LC-MS-based proteomics. maSigPro has been used to analyze data from intrinsically dynamic processes such as the spatial differentiation in fungi (Levin et al. 2007), and plant development (Pascual et al. 2009; Wong et al. 2009a; Wong et al. 2009b), as well as periodic responses such as the rhythmically expressed genes in mouse distal colon (Hoogerwerf et al. 2008).

3.7.1.2.3 ANOVA-SCA

ANOVA-SCA (or ASCA for short) is considered a combination of a statistical method (ANalysis Of VAriance, ANOVA) and a dimensionality reduction approach (Simultaneous Component Analysis, SCA) (Jansen et al. 2005; Smilde et al. 2008; Smilde et al. 2005). ANOVA-SCA is particularly useful when two or more quantitative variables are involved, such as time and dose. In the first step, an ANOVA model is applied for each gene expression measure to separate the variability caused by these two different variables. The model parameters obtained for all genes under each experimental condition are subsequently organized into a matrix form. The second step involves applying principle component analysis simultaneously on all matrices obtained under all experimental conditions. A number of constraints can be further imposed such that the resulting matrices are mutually independent. Such constraints on orthogonality enable

the ASCA model parameters to be estimated independently by solving a simple least-square optimization problem. Statistical significance of these experimental variables and their interactions can be further inferred using a permutation approach (Vis et al. 2007). In particular, all experimental conditions are permutated to obtain a no-effect distribution, thus providing a baseline to conclude whether the observed effect is indeed significant.

One of the earliest applications of ASCA resides in analyzing a metabolomics experiment wherein the effects of time and vitamin C dose on the NMR spectra of guinea pig urine samples were delineated (Smilde et al. 2005). Individual variations caused by time and doxorubicin dose on metabolite MS profiles were also uncovered using ASCA in a toxicology study on rats (Wang et al. 2009a). Given the intrinsic generalizability of ASCA, it is not surprising to find this approach extended into discovery of kinetically differentially expressed genes (Nueda et al. 2007). Two statistics – SPE (Squared Prediction Error) and leverage – were proposed to evaluate the goodness of fit of the ASCA model, and the degree of agreement by which a gene profile follows the main expression patterns, respectively. This adapted version of the original algorithm, *ASCA-genes*, has been implemented in the R language. Furthermore, *ASCA-fun* was devised to perform functional analysis on time-series microarray data (Nueda et al. 2009). In this method, genes ranked according to their correlation to the principal time components identified by ASCA were used to assess functional enrichment in the dataset following Gene Set Analysis (GSA) procedures.

3.7.1.2.4 Bayesian Approaches

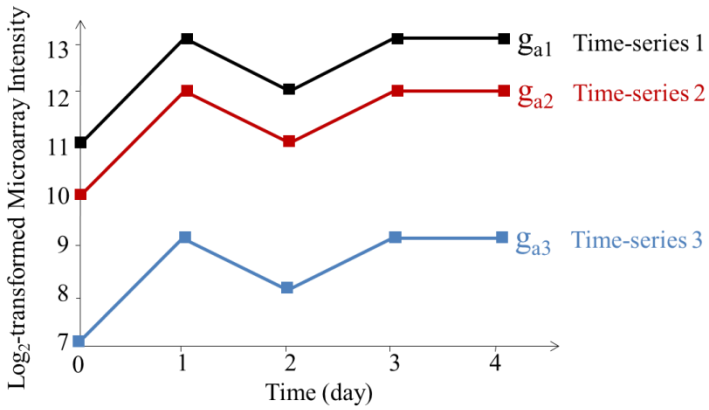
A multivariate empirical Bayes model was applied to time-series microarray data by Tai and Speed (Tai and Speed 2006). The algorithm, implemented in the R package *timecourse*, however, requires replicates of the full time-series. This algorithm calculates multivariate versions of the log-odds, or *B*-statistic (*MB*-statistic), and the Hotelling statistic (T^2). When the numbers of replicates are the same for all genes, the *MB*-statistic is equivalent to the T^2-statistic. The algorithm can be used in one-treatment problems and multi-treatment problems. Although this method ranks the genes, it does not provide a significance cutoff.

A fully Bayesian approach for microarray analysis was implemented in clustering (Heard et al. 2006) and later for the analysis of time series (Angelini et al. 2007). This fully Bayesian approach can handle short series, non-uniform sampling and missing data and does take into consideration the temporal structure of the time series. Gene expression profiles are modeled with Legendre or Fourier polynomials, and the coefficients and the degrees of these polynomials are estimated using a Bayesian approach. The differentially expressed genes identified in this Bayesian multiple-testing procedure are ranked, and their expression profiles are estimated. This estimation allows the visualization of each gene expression profile as a single smooth curve.

The fully Bayesian approach was demonstrated when analyzing the time series obtained by stimulating human breast cancer cells with estradiol after different time periods. The algorithm is implemented in the Bayesian user-friendly software for Analyzing Time Series (BATS) (Angelini et al. 2008), a graphic user interface written in Matlab. The BATS package requires 5-6 time points and replicates are recommended but not required. At the moment, however, BATS can only handle one-treatment time series. Its extension to multiple time series is under development.

3.7.2 Calculation of Distances between Gene Expression Profiles

Just as in the calculation of the geometrical distance between any two vectors, a distance value can be computed to quantitatively describe the difference between two expression profiles of the same gene. By condensing all distance measures between the corresponding time points, the comparison of these two profiles is reduced into a single number. Two frequently used metrics are Euclidean distance and Pearson's correlation (Figure 7).



Euclidean distance

$$Eucl(g_{a1}, g_{a2}) = \sqrt{\sum_{i=0}^{4}(g_{a1,i} - g_{a2,i})^2} = \sqrt{(11-10)^2 + (13-12)^2 + (12-11)^2 + (13-12)^2 + (13-12)^2} = 2.2$$

$$Eucl(g_{a1}, g_{a3}) = \sqrt{\sum_{i=0}^{4}(g_{a1,i} - g_{a3,i})^2} = \sqrt{(11-7)^2 + (13-9)^2 + (12-8)^2 + (13-9)^2 + (13-9)^2} = 8.9$$

Pearson correlation

$$Corr(g_{a1}, g_{a2}) = \frac{\sum_{i=0}^{4}(g_{a1,i} - \overline{g_{a1}})(g_{a2,i} - \overline{g_{a2}})}{\sqrt{\sum_{i=0}^{4}(g_{a1,i} - \overline{g_{a1}})^2 (g_{a2,i} - \overline{g_{a2}})^2}} = 1 \rightarrow \text{Pearson distance } (g_{a1}, g_{a2}) = 0$$

$$Corr(g_{a1}, g_{a3}) = \frac{\sum_{i=0}^{4}(g_{a1,i} - \overline{g_{a1}})(g_{a3,i} - \overline{g_{a3}})}{\sqrt{\sum_{i=0}^{4}(g_{a1,i} - \overline{g_{a1}})^2 (g_{a3,i} - \overline{g_{a3}})^2}} = 1 \rightarrow \text{Pearson distance } (g_{a1}, g_{a3}) = 0$$

Figure 7: Calculation of distance between the expression profiles of a gene in two series.

The distance between the expression profiles of gene g_a in three series can be measured using different metrics. The Euclidean metric quantifies the absolute geometric distance between the profiles, whereas the Pearson metric evaluates the correlation of trends in expression. Thus even though the Euclidean distance of g_a between series 1 and series 3 (Eucl (g_{a1}, g_{a3})) is much higher than that between series 1 and series 2 (Eucl (g_{a1}, g_{a2})), their Pearson distances (Corr(g_{a1}, g_{a2}) and Corr(g_{a1}, g_{a3})) are indeed the same.

Euclidean distance, also known as *L-2 norm*, assesses the absolute difference between two time profiles. As a result, genes with the highest Euclidean distance between two treatments are often the ones with a high expression level, and are most likely to be identified as differentially expressed despite having similar expression trends in these treatments. Gene expression data can thereby be mean-centered or *z*-transformed to alleviate the dominance of these high-abundance transcripts. On the other hand, Pearson's correlation quantifies the overall similarity between the two trends regardless of the absolute values of gene expression. Small fluctuations in gene expression between

low-abundance transcripts can thus be manifested as being markedly different since only expression trend is considered.

The choice of distance metric therefore depends on the question being asked. If the absolute values of expression measures are critical, the Euclidean metric is often preferred. Alternatively, the Pearson's correlation coefficient is a more suitable similarity measure if the overall trend of expression is pertinent to the analysis. A combination of both metrics is therefore recommended to integrate the differences in absolute expression magnitude and expression trend.

Following selection of a proper metric and distance calculation, a distribution of this representative difference can be plotted, and a threshold is often set to declare differential expression. Genes having distance measures between the expression profiles in two treatments above a certain threshold are considered differentially expressed. Manual inspection of gene expression profiles is often recommended to confirm the differential expression. In addition, if both treatments were replicated, a statistic can be derived by permuting replicated samples between the two treatments. An average distance over all permutations is calculated, and compared to the actual distance to infer a statistical significance level. Yet optimizing the difference threshold between the average and the actual distance can be challenging.

This approach was used in a number of studies conducted in *Streptomyces coelicolor*. Genes involved in regulatory circuits related to antibiotic production were identified using Euclidean distance as a criterion for differential expression (Mehra et al. 2006). Euclidean distance was also used in conjunction with principal component analysis (PCA) to reveal genes kinetically perturbed when the *Streptomyces coelicolor* sigma-like protein AfsS was disrupted (Lian et al. 2008). In a recent study, more than 900 genes were identified as differentially expressed in an antibody-producing CHO cell line between the butyrate-treated 33°C culture and the non-treated culture (Kantardjieff et al. 2010).

3.8 PROFILE PATTERN RECOGNITION

Microarray data, with their large size and high-dimensionality, are inherently complex. Compared to the number of genes (i.e., dimensionality), the number of samples is almost always small, making it difficult to find an answer to the question being asked. Often, an objective in a microarray experiment is to identify genes with a certain profile or pattern. Sometimes, however, which patterns are present in the data are not even known. In order to identify patterns that exist in the data, two types of techniques can be used: unsupervised and supervised algorithms.

3.8.1 Unsupervised Classification Methods

Unsupervised pattern recognition consists of organizing data based on the properties of data themselves without reference to additional information (Gollub et al. 2006). Mathematical algorithms determine the search for natural patterns existing in the data (Morrison and Ellis 2003). The goal of unsupervised pattern recognition is to identify small subsets of genes that display similar expression patterns (Boutros and Okey 2005). Instead of clustering genes, clustering samples based on their expression profiles can also be a goal in clustering analysis. In this case samples with similar expression profiles might help identifying groups, or labels, that can be given to those samples.

Although the term unsupervised pattern recognition is commonly used as a synonym for "clustering", it actually encompasses other techniques, such as non-negative matrix factorization (NMF) and principal component analysis (PCA).

3.8.1.1 Dimensionality Reduction Techniques

Because microarray data often are obtained from only a small number of samples and entail thousands of genes, dimensionality reduction can be helpful for visualization, clustering, and classification. When transcriptome data are represented as an n by m matrix, in which n is the number of genes and m the number of samples ($n >> m$), dimensionality reduction techniques can be used to identify a smaller number k of principle gene expression patterns (Figure 8). This can be done by factorizing the original gene expression matrix (A) into two sub-matrices: one containing eigenarrays (W) and the other containing k eigengenes (H). The expression level of each gene in

these *m* samples can be represented as a linear combination of the *k* eigengenes. Similarly, the overall expression pattern in each sample can be represented as a linear combination of the *k* eigenarrays.

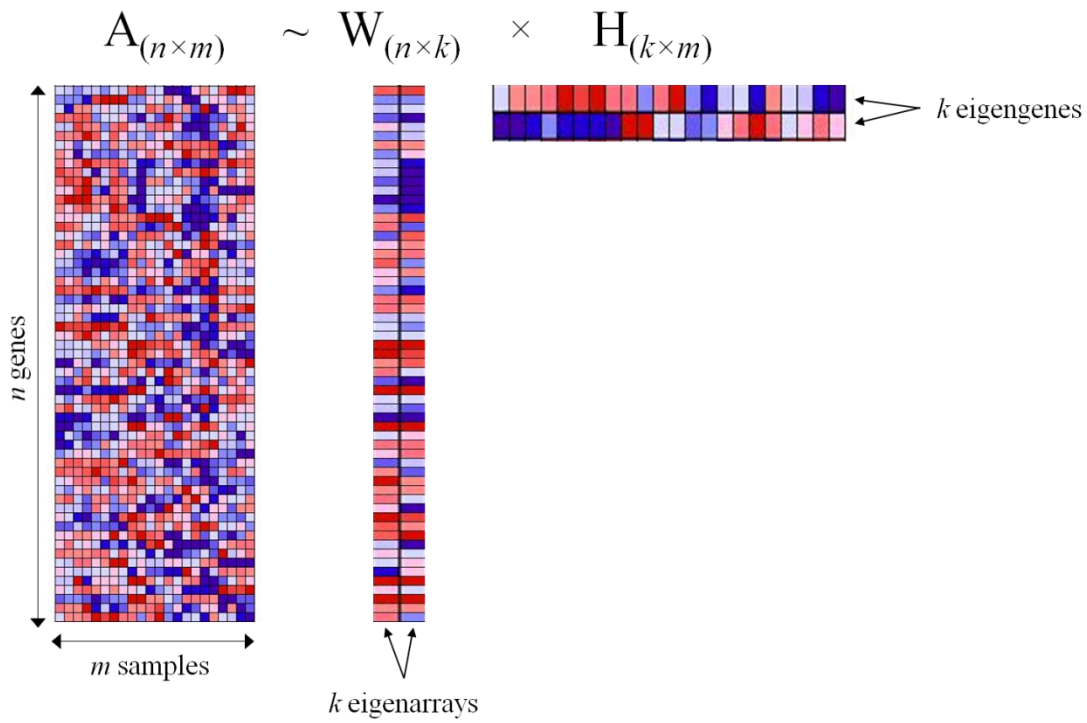


Figure 8: Matrix factorization in dimensionality reduction techniques: NMF and PCA.

Microarray data are organized into a matrix (A) with each row representing the expression levels of a gene in *m* samples. This original matrix can be decomposed into two sub-matrices: one containing *k* eigenarrays (W), and the other containing *k* eigengenes (H). The expression levels of each gene in these *m* samples can be represented as a weighted combination of the *k* eigengenes. Similarly, the overall gene expression pattern in each sample can be represented as a weighted combination of *k* eigenarrays.

In principal component analysis (PCA), the data are transformed into a new set of variables called principal components (PCs). The principal components are uncorrelated, and furthermore, they are ranked so that the first PCs contain most of the variation present in all of the original variables (Jolliffe 2005). Since the first few PCs capture most of the variation in the original data, it is customary to use only the first few PCs (Yeung and Ruzzo 2001). When the data are projected along the first few PCs (most commonly the first two or three), in many cases it is possible to identify groups.

In PCA, the gene expression values can be reconstructed by a weighted sum of the eigengenes, however there is no restriction on the sign of the weights. This can cause some variability due to cancellations, if eigengenes with both negative and positive weights are added. In a similar technique, Non-negative Matrix Factorization (NMF), the coefficients are forced to be non-negative, which ensures that the contributions from principal gene expression patterns are positive and thus additive (Brunet et al. 2004; Lee and Seung 1999).

Both techniques, PCA and NMF, have been used in the identification of biomarkers, for example see (Schachtner et al. 2007). PCA has been used to characterize the gene expression of stem cells in different phases (Aiba et al. 2006) and different types of stem cells (Ulloa-Montoya et al. 2007). As NMF has been found superior to PCA in reducing microarray data (Liu et al. 2008), it has been used more extensively in the identification of cancer molecular patterns for gene expression data (Brunet et al. 2004; Frigyesi and Hoglund 2008; Han 2008).

3.8.1.2 Clustering

Clustering is one of the most widespread tools for grouping transcripts in microarray data. The concept of clustering is based on the simple idea of grouping similar objects. The goal is to maximize the similarity between objects in the same cluster, and minimize the similarity of objects in different clusters. How similarity is measured is thus a key part of clustering algorithms. In the case of microarray data, the expression profile of a gene, made up by the different samples, is seen as a series of coordinates that define a vector (Sherlock 2000). Distance metrics can thus compare the similarity of the direction and/or magnitude of two or more vectors.

Traditional clustering algorithms have existed since the 1950s and have been applied to a number of problems, including image analysis, marketing, document classification, and population studies. These traditional algorithms have also been used to cluster transcriptome data. In addition, specialized clustering algorithms have been developed for time-series data.

3.8.1.3 Clustering of Static Samples

In the case of static sampling, transcriptome data can be represented as a matrix, with each row representing a gene, and each column representing a single condition. The data can thus be represented as vectors and the distance between these vectors can be determined. Note that there are two ways to organize the data. One is to take the expression value of each gene across different samples as a vector. The other one is to take the expression of all genes in a sample as a vector. Clustering can thus be used to find genes behaving similarly in different samples or samples which are "similar" in overall gene expression. In the following section, all examples are illustrated as clustering genes with similar transcriptional behaviors in different samples. The alternative of classifying samples based on their overall gene expression data is demonstrated in the supervised classification topic.

A distance measure (such as Euclidean, Manhattan, Chebyshev, Mahalanobis, Pearson, cosine, Spearman, or Kendall) is used to assess similarity and the data are then organized into clusters according to clustering rules. These clusters can be of fixed size, (i.e. the number of clusters determined *a priori*), or natural clusters discovered in the data. The most commonly used clustering algorithms broadly correspond to two categories: hierarchical clustering and partitional clustering.

Hierarchical clustering can be bottom-up, starting with single-gene clusters and joining the most similar clusters until a single cluster with all genes is obtained; or top-down, starting with all genes in a single cluster and dividing them into smaller clusters (Nugent and Meila 2010). In both cases, the result is represented as a hierarchical tree, or dendogram. Most commonly, the bottom-up approach is used (Figure 9). Initially, two closest genes (1 and 2; 3 and 4) are joined using one of the distance metrics. In the next iteration, a linkage or amalgamation rule is needed to join these multiple-gene clusters (Frades and Matthiesen 2010). This rule can be single-linkage, complete-linkage, or average linkage. In single linkage (aka nearest neighbor), the similarity of these two clusters is the shortest distance of all pair-wise comparisons of the genes in one cluster to the other; in this example, the distance between gene 1 and gene 3. In complete linkage (aka furthest neighbor), the similarity of these two clusters is defined as the largest

distance of these pair-wise comparisons; in this case, the distance between gene 2 and gene 4. In average linkage, the distance between these two clusters is that between their centroids (Gollub et al. 2006). In this instance, the centroid of the first cluster is a hypothetical gene "in the middle of" gene 1 and gene 2, and thus its expression level is taken as the average expression level of these two genes.

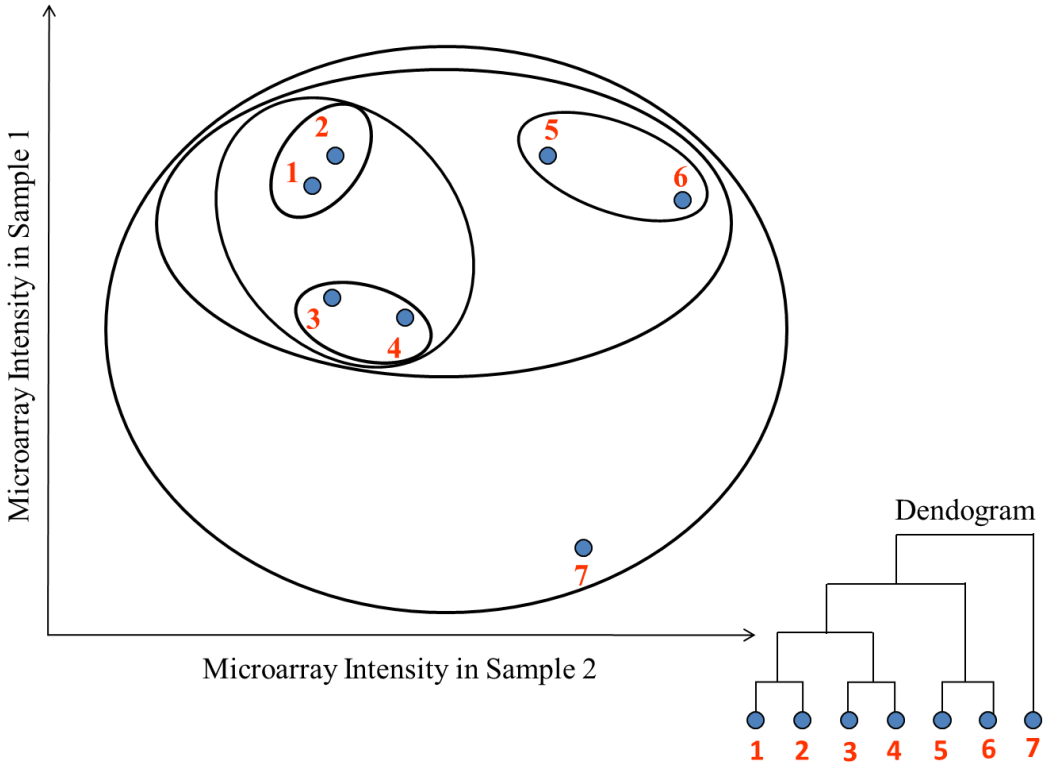


Figure 9: Simple illustration of hierarchical clustering.

The algorithm starts with each gene belonging to its own cluster, followed by joining the two closest genes: 1 and 2. Subsequently, individual genes or multi-gene clusters are joined using single linkage, complete linkage, or average linkage. In this case, single linkage is used, i.e., the distance between two clusters is taken as the shortest distance between any two members of the clusters. Thus the distance between cluster 1-2 and cluster 3-4 is the distance between genes 1 and 3. The two closest clusters are joined accordingly, in this case cluster 1-2 and cluster 3-4. This grouping is continued until all genes are joined into one cluster, and the whole process can be visualized as a dendogram.

Hierarchical clustering has been used extensively to compare cell types and tissues, including diseased vs. healthy cells, and drug effects, for example (Ambrosi et al. 2007; Anichini et al. 2003; Fortier et al. 2010; Secco et al. 2009; Vega et al. 2006). Hierarchical clustering has also been used to classify proteomic profiles of serum,

plasma, and modified media supplements used in cell culture (Ayache et al. 2006), and metabolomic profiles of extracellular metabolites in recombinant CHO fed-batch cultures (Chong et al. 2009).

In partitional clustering, data points are separated into a pre-defined number of clusters. In the first step of these iterative algorithms, data points are randomly assigned to clusters. The distance between individual data points and the clusters is then calculated and used to reassign the data points to the cluster to which they are closest. This process continues until all data points are assigned to the closest cluster (De Bruyne et al. 2005). *K*-means clustering, Self-Organizing Maps (SOM), and Fuzzy C-means (FCM) clustering are among the best known clustering algorithms in this category. One limitation of these algorithms is that the number of clusters has to be fixed from the beginning, and thus the results are dependent on it (Dopazo et al. 2001).

In *k*-means clustering (Everitt 1974b), *k* is the number of clusters and is a required input. *k* random points are used as cluster centers (or means) at initialization. All data points are assigned to these initial clusters by finding the one with the closest distance. In iterative steps, the mean of each cluster is recalculated and the data points reassigned to new clusters (Do and Choi 2008). This process continues until the assignment does not change considerably. As the value of *k* greatly influences the final outcome, several algorithms include a procedure to determine the best *k*. *k*-means clustering has been used to analyze transcriptome data of cancer cells (Liu et al. 2010b), and stem cells (Ulloa-Montoya et al. 2007; Way et al. 2009) among others.

Similar to *k*-means clustering, in the case of SOMs (Kohonen 2001), the number of clusters is also a required input. In addition, their geometry must be specified (grid size). Thus not only the number of clusters but also their geometry has an effect on the final clustering result. A seed vector is first assigned to each cluster, and data assigned to these clusters in an iterative process. In each iteration, gene expression data randomly selected is compared to the seed vectors. The gene is assigned to the cluster that has the more similar seed vector. The value of the seed vector is updated, so that it is more similar to the expression of the gene used in the comparison. Because the cluster centers are part of a grid, the values of the other seed vectors are also modified although to a

lower extent. SOMs have been used to analyze monolayers of cultured rat hepatocytes (Baker et al. 2001), to study hematopoietic differentiation (Tamayo et al. 1999), to investigate the saline osmotic tolerance in yeast (Pandey et al. 2007), and to investigate hepatic differentiation (Li et al. 2007b), among others.

Whereas *k*-means and SOM assign each gene to a single cluster (hard clustering), FCM (Bezdek 1981) links each gene to all clusters using a series of values. Values close to 1 indicate strong association to a cluster, and values close to 0 indicate absence of association. These indexes define the membership of each gene with respect to all clusters (Dembele and Kastner 2003). In addition to the number of clusters, the fuzziness parameter is also a required input. Kim et al. (Kim et al. 2006) have reported that the fuzziness parameter is sensitive to the normalization method used, and thus the clustering results vary with the normalization method. Recently, a method for the determination of the optimal parameters for FCM has been proposed (Schwammle and Jensen 2010). FCM has been used to analyze gene expression profiles in high-grade gliomas (Czernicki et al. 2007) and in tumor sample classification (Wang et al. 2003).

3.8.1.4 Clustering of Dynamic Samples

Clustering algorithms such as hierarchical clustering, *k*-means, and SOM, are also commonly used to analyze time-series data. However, these algorithms do not take into account the sequential aspect of time-series data (Tchagang et al. 2009). Thus clustering of time series requires specialized algorithms. Some of the specialized algorithms require long series (> 10 time points), whereas others have been developed specifically for short time series.

B-splines (Bar-Joseph et al. 2003; Gaffney and Smyth 2005; Luan and Li 2003), linear splines (De Hoon et al. 2002), ordered restricted inference (Peddada et al. 2003), hidden Markov models (Schliep et al. 2003), and gene expression dynamics using regression (Ramoni et al. 2002) are examples of clustering algorithms that can be used for long time-series data. Fuzzy C-Varieties with Transitional State Discrimination preclustering (FCV-TSD) (Moller-Levet et al. 2003), ASTRO and MiMeSR (Tchagang et

al. 2009), and Short Time-series Expression Miner (STEM) (Ernst and Bar-Joseph 2006) are examples of clustering algorithms developed specifically for short time series data.

STEM selects a set of potential expression profiles, each representing a unique pattern. Each gene is then assigned to the profile that best represents it. The significance of each profile is determined using hypothesis testing. The number of genes assigned to each profile under the true ordering is compared to the average number of genes assigned to each profile when permutated data are used. The significant profiles can then be analyzed independently or grouped into clusters. STEM has been used to cluster time-course microarray data collected in the study of egg development in *Drosophila melanogaster* (Baker and Russell 2009), salt stress in *Medicago truncatula* (Lievens et al. 2009), and muscle differentiation (Ozbudak et al. 2010).

Biclustering takes clustering algorithms a step further. It consists of simultaneous clustering of both genes (rows) and conditions (columns) (Cheng and Church 2000). The goal in biclustering is to find submatrices (Madeira and Oliveira 2004), that is, to identify subgroups of genes and/or subgroups of conditions with highly correlated behaviors. Thus biclustering can find correlations in certain datasets where other algorithms cannot. Biclusters can be of constant row, constant column, or both constant row and column.

Among the software that can perform biclustering are Gene Expression Mining Server (GEMS) (Wu and Kasif 2005), Expression Analyzer and DisplayER (EXPANDER) (Shamir et al. 2005), Phase-shifted Analysis of Gene Expression (PAGE) (Leung and Bushel 2006), Biclustering Gene Expression Time Series (BIGGEsTS) (Goncalves et al. 2009), Biclustering algorithm and Visualization (BiVisu) (Cheng et al. 2007), and Biclustering Analysis Toolbox (BicAT) (Barkow et al. 2006), which integrates several biclustering algorithms.

3.8.2 Unsupervised Classification Methods

Unsupervised classification methods are used for the identification of naturally existing clusters within the data. Supervised approaches, on the other hand, are designed to address the following question – Given a set of samples categorized into pre-defined groups (training set); can we use the gene expression data of these samples to construct a

rule, or a function, to differentiate these groups? This also implies the ability to use this rule for classification of new, uncategorized samples (test set) based on their expression data.

Since the classification rule is built upon the training set, it may fit this dataset "too well" and thus have poor performance on unclassified samples in the test set (Figure 10a). In this example, the high producer clones (blue circles) and the low producer clones (black squares) can be simply separated by a linear model (solid line), allowing several samples to be misclassified (outliers). Yet the model can become over-complicated (dash line) when trying to classify correctly all outliers and thus often results in higher error rate in classifying regular samples. This is known as "overfitting", and ideally should be assessed using an independent test set. However, in situations where acquiring additional data is expensive or not feasible, various cross-validation schemes can be used. The leave-one-out scheme allocates one sample for testing whereas the rest are used to train the classification model. In the hold-out scheme, the data are split into two equal sets – one is used for training, and the other for testing. Another frequently used method is the *k*-fold cross-validation, in which the data are divided into *k* sets – the first (*k*-1) sets are used for training, and the last one for testing (Figure 10b). This process is repeated until all data have been used for testing. Commonly used supervised classifiers for gene expression data include *k*-Nearest Neighbors (KNNs), decision trees, Artificial Neural Networks (ANNs), and Support Vector Machines (SVMs). These algorithms have been implemented in several code libraries and various downloadable packages in Matlab and R.

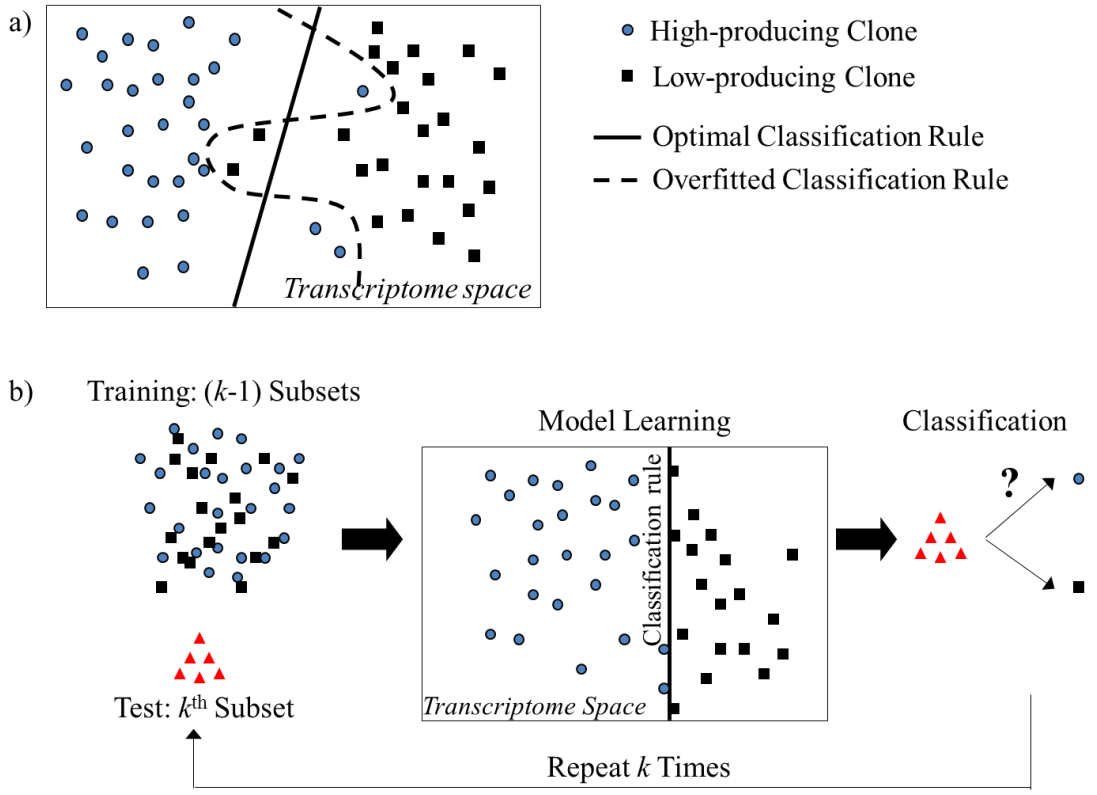


Figure 10: Overfitting of training data and *k*-fold cross validation scheme.

a) High producer clones (blue circles) and low producer clones (black squares) can be separated using a linear model (solid line) with a few outliers. Yet the model can become complicated (dash line) when all outliers are taken into consideration. This overfitted model will have high error rate when classifying new samples.

b) The data are split into *k* subsets: (*k*-1) subsets are used for training the model, and testing is performed on the k^{th} subset. This process is repeated *k* times until all data have been used for testing.

3.8.2.1 K-nearest Neighbors (KNNs)

KNN is among the simplest and most fundamental of the classification methods, and is often the first choice when prior knowledge about the dataset is minimal. Given that a set of samples has been classified into different groups, a new sample will be assigned into the group whose members constitute the majority in the neighborhood of the sample (Cover and Hart 1967; Tan et al. 2005). The choice of distance metric thus becomes vital in this case – a new sample can be assigned to a different group when a different distance metric is applied. In addition, if a certain group is dominated in size compared to the others, a bias in assigning new samples into that group is likely to occur. One way to circumvent this problem involves giving each "neighbor" a weight inversely

proportional to its distance to the new sample. Furthermore, the distance threshold and the number of "neighboring" samples *k* also have an effect on the final classification, and thus should be optimized using cross-validation.

In the FDA MicroArray Quality Control (MAQC) project, a KNN data analysis protocol was developed to predict the clinical outcome of about 500 new neuroblastoma patients (Parry et al. 2010). These KNN models were built using a large gene expression dataset obtained from approximately 700 breast cancer, neuroblastoma, and multiple myeloma samples. In another example, gene expression signatures from 4413 probes in 37 colorectal cancer samples were also used to train a KNN model which was further validated using a leave-one-out scheme (Laiho et al. 2006). This model successfully classified these samples into serrated and conventional colorectal cancer samples using the expression data of 10 genes.

3.8.2.2 Decision Trees

Decision trees are built using an iterative scheme wherein a question about the gene expression signatures of the training samples is posed at each node (Breiman et al. 1984; Kingsford and Salzberg 2008; Quinlan 1993; Tan et al. 2005). The entire tree is obtained by repeated splitting of those samples into two or multiple descendant subsets. The training samples will guide the choice of splitting rules such that each terminal node of the tree, i.e., leaf, is assigned a group label. Thus decision trees are often more interpretable than other classifiers, and naturally support multiple-group assignment. Furthermore, multiple decision trees can be combined into an ensemble, e.g., random forest, to increase the classification accuracy (Breiman 2001; Tong et al. 2003). When applying decision trees, it is critical to control the complexity of the tree, i.e., avoid overfitting the training data. In addition to using cross-validation, one can also prune the tree by collapsing several internal nodes into one leaf, or stop branching the tree when there is no substantial improvement in the homogeneity of the final group assignment.

Several decision tree algorithms were applied to 869 genes differentially expressed in earthworms in response to explosive compounds TNT or RDX (Li et al. 2010). Over 350 genes were subsequently selected by these algorithms as classifiers, and

ranked according to their significance in the assembled tree. In another application, hierarchical clustering results of gene expression data from three different cohorts of 481 breast cancer samples were further analyzed using decision trees (Ihnen et al. 2010). Four groups with different expression levels of osteopontin (OPN), activated leukocyte cell adhesion molecule (ALCAM), human epidermal growth factor 2 (HER2), and estrogen receptor (ER) were found. Patients with high OPN and low ER, HER2 and ALCAM were placed in a particularly high risk group.

3.8.2.3 Artificial Neural Networks (ANNs)

ANNs were developed based on the computation principles occurring in the network of neurons within the human brain (Krogh 2008; Minsky and Papert 1969; Tan et al. 2005). An ANN model can be considered as an assembly of interconnected nodes in which all input sources, in this case the expression values of all genes on the array, are weighted and combined. This weighted average is compared to a threshold, yielding an output value based on a step function. If the average exceeds the threshold, the output value will be one, corresponding to one group; otherwise it is zero, which corresponds to the other group. During the training process, the weighting factors and the threshold can be estimated iteratively, and a linear decision boundary (i.e., separating hyper-plane) can be obtained. Yet when the data are not linearly separable, hidden layers of intermediate nodes can be added to the network. A partial classification is performed at each layer, and assembled to achieve the final classification at the output node. Furthermore, alternative functions such as sigmoid or linear models can be utilized in place of the simple step function in these feed-forward neural networks.

Using gene expression data obtained from 63 training samples of small, round blue cell tumors (SRBCTs), 3750 ANNs have been constructed and cross-validated (Khan et al. 2001). Without over-fitting, these models successfully classified the samples into four diagnostic categories of tumors. ANNs have also proven efficient in tracking transcriptional changes responsible for progression from the chronic stage to a highly aggressive acute stage of adult T-cell leukemia (ATL) (Choi et al. 2006). Using gene expression data from more than 44,000 probe sets and 10-fold cross validation on 37

samples, 44 “predictor” genes could be identified, offering the possibility to diagnose different ATL stages.

3.8.2.4 Support Vector Machines (SVMs)

In binary SVM, two groups (for example, high producer clones and low producer clones) are separated in such a way that the distance between the training samples and the decision boundary is maximized (Boser et al. 1992; Noble 2006; Tan et al. 2005) (Figure 11). This optimization process results in the construction of a separating hyper-plane, e.g., a linear line in 2-dimensional space, which maximizes the margin between the two groups. In several cases where the samples are not linearly separable in the original space, a kernel function can be chosen to transform the data to a higher-dimensional space in which a “linear” hyper-plane can be found. Furthermore, a few anomalous samples are often allowed to be misclassified to achieve a larger margin. Thus a cost function has to be selected and optimized such that the size of this “soft” margin is balanced with the allowable degree of hyper-plane violation.

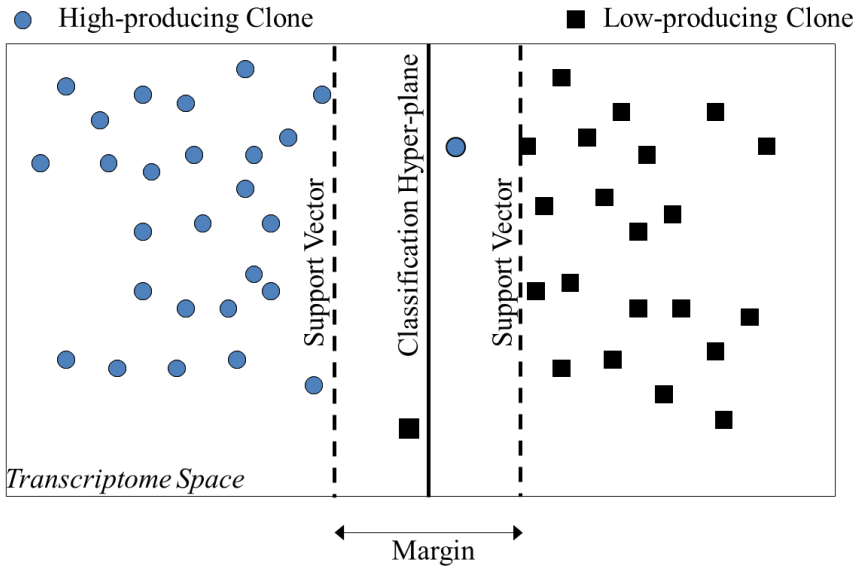


Figure 11: Support Vector Machines (SVMs) with soft margin.

Binary SVM algorithms search for a separation hyperplane that maximizes the margin (or distance) between two groups: in this case, high producer clones and low producer clones. Samples on the margins are referred to as “support vectors”. A few samples can be misclassified in order to obtain a maximal margin (“soft margin”).

Gene expression data from 97,802 clones were used to construct several SVM models using the simple dot-product kernel and validated through the leave-one-out scheme (Furey et al. 2000). Thirty-one human tissue samples were successfully classified by these models into cancerous ovarian and normal tissues. Interestingly, an SVM model was also built using gene expression profiles from seven high and four low recombinant IgG-producing NS0 cell lines. Through the leave-one-out cross-validation process, the transcriptomic differences between these high and low producers were indeed highlighted, supporting the molecular basis of productivity trait (Charaniya et al. 2009).

3.9 PATHWAY ANALYSIS

Microarray analysis results in a list of differentially expressed genes or genes with a dynamic trend over time. It is possible that the transcriptional changes seen on those genes might not be independent, but rather have occurred in a coordinated manner. Thus understanding the physiological relevance of these changes requires analysis in a biological context, beyond what differential expression analysis can determine. Furthermore, examining genes in each pathway as a whole allows one to detect subtle, yet consistent, transcriptional changes that would otherwise be neglected by differential gene expression analysis.

Pathway analysis involves mapping the list of differentially expressed genes onto known pathways in order to elucidate a whole chain of events which might have occurred during the experiment. Depending on the microarray platform, probe IDs can be linked to different sources of annotation, for instance, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Map Annotator and Pathway Profiler (GenMAPP). This retrieval of pathway information allows all differentially expressed genes in a certain pathway to be highlighted. Yet statistical tests need to be performed to confirm whether the entire pathway is indeed enriched or under-represented rather than occurring by random chance. A number of methods and software have been developed to assess the statistical significance of this functional enrichment/under-representation, including Ingenuity's IPA, GeneGo's MetaCore, GenMAPP's MAPPFinder, Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005), and

Gene Set Analysis (GSA) (Efron and Tibshirani 2007). Those methods differ in the calculation of the enrichment score and the corresponding significance level, usually *p*-value or FDR. For illustrative purposes, two representative methods, MAPPFinder and GSEA, are described in the following section.

3.9.1 MAPPFinder

In order to assess the degree of enrichment for each pathway (or gene set), MAPPFinder calculates a *z*-score using the number of differentially expressed genes in the set, the number of genes in the set, the number of differentially expressed genes in total, and the total number of genes on the array (Dahlquist 2002; Dahlquist et al. 2002; Doniger et al. 2003). A high positive *z*-score indicates that the pathway of interest is significantly enriched, and an extreme negative *z*-score suggests that it is under-represented. Furthermore, if a *p*-value is desired, a *z*-score of 1.96 or -1.96 can be converted to a *p*-value of 0.05 given that the data strictly follows a hyper-geometric distribution. It is important to note that similar to several other pathway analysis tools, MAPPFinder also requires a pre-defined list of differentially expressed genes. This is sometimes challenging since the list can vary considerably depending on the selected fold-change and the *p*-value cutoff.

Prickett et al. have demonstrated the use of MAPPFinder in uncovering several immune-system pathways affected in chicken infected with a protozoan parasite (Prickett and Watson 2009). About 1,175 genes, accounting for about 10% of total unique Ensembl genes present on the array, were mapped to 85 inferred chicken pathways in GenMAPP, 18 of which were either up- or down-regulated at a *p*-value cut-off of 0.05. In another study, functional enrichment information obtained from MAPPFinder was linked automatically to the original gene expression data to calculate the average intensity or ratio of all differentially expressed genes in each pathway (Yu et al. 2006). This quantitative evaluation of dose- and time-dependent microarray data in rats exposed to toxicants thus allows one to calculate an effective dose (ED_{50}) for each pathway, which plays an important role in risk assessment.

3.9.2 Gene Set Enrichment Analysis (GSEA)

GSEA is a powerful tool for pathway analysis which calculates gene set enrichment using all genes present on the array instead of a pre-defined set of differentially expressed genes (Gene Set Enrichment Analysis ; Mootha et al. 2003; Subramanian et al. 2005). An ordered list is first generated by ranking all genes in the dataset based on their signal-to-noise ratio (Figure 12).

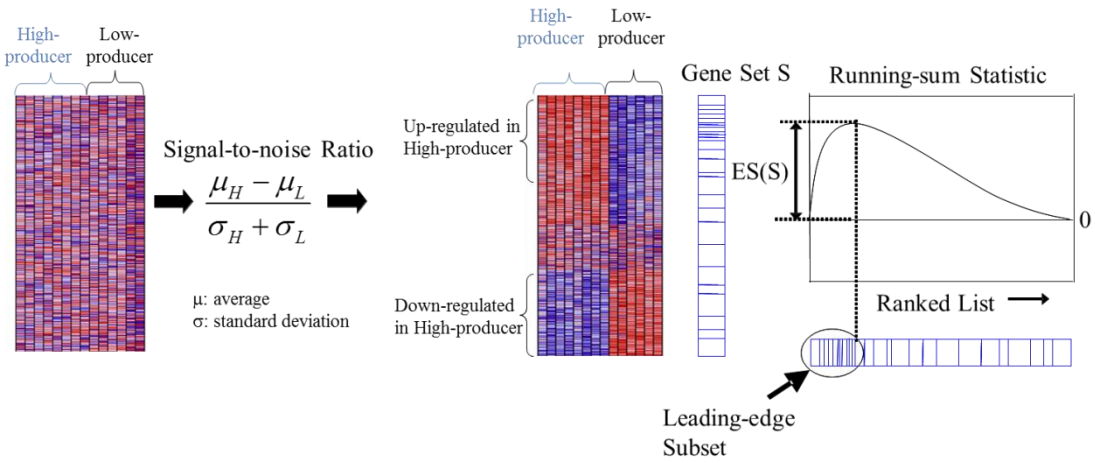


Figure 12: Gene Set Enrichment Analysis (GSEA).

Genes are ranked based on their signal-to-noise ratios to create an ordered list. A running sum statistic is calculated by walking down this list. If the encountered gene is part of the gene set of interest, the running sum statistic is increased; otherwise, it is decreased. The enrichment score of each gene set is chosen as the maximum deviation of this statistic from zero. Genes with key contributions to the enrichment of the gene set are listed in the leading edge subset.

This ratio is often the quotient between the difference in average expression levels and the overall variability of measurement. In the second step, a running-sum statistic is measured for each pathway (or gene set) by travelling down the ordered list. If the encountered gene is a part of the gene set of interest, the statistic is increased; otherwise it is decreased. The magnitude of this change is set to be proportional to the signal-to-noise of that gene and the size of the gene set it belongs to. The maximum deviation from zero of the running-sum statistic is chosen as the enrichment score (ES), and an associated statistical significance (*p*-value) can be calculated using a permutation scheme. Concurrently, a leading-edge subset of genes, which are key contributors to enrichment of the functional class represented by the gene set, can also be exported.

Deregulated functional categories in Ewing's sarcoma family tumors (ESFT) cell lines under hypoxia were identified by applying GSEA with three different gene sets (Aryee et al. 2010). Hypoxia-related functions such as angiogenesis, vasculature development, and glucose metabolism were shown to be up-regulated under hypoxic conditions. GSEA was also used alongside other pathway analysis tools to investigate the biological relevance of transcriptional differences between neurofibromatosis type 1 (NF1)-haploinsufficient lymphoblastoid cell lines (LCLs) and mouse B lymphocytes (Pemov et al. 2010). Despite the modest changes in gene expression detected using *t*-test, several pathways were revealed to experience perturbations including cell cycle, DNA replication and repair, transcription and translation, and immune response.

3.10 NETWORK RECONSTRUCTION

Gene network inference attempts to reconstruct gene networks reflecting their interactions from high-throughput data, especially microarray data. Network reconstruction is a challenging task as gene interactions are dynamic and membership of particular elements in a network is not always permanent. In this regard, the use of microarray data compiled under a wide range of conditions, or from a variety of mutants, can help unveil interactions. Also, time-series microarray data are of particular relevance in reverse engineering regulatory networks. In addition to algorithms for constructing regulatory networks using static gene expression data, special algorithms have also been developed for data obtained from time-series microarrays.

3.10.1 Network Reconstruction from Static Gene Expression Data

3.10.1.1 Information Theoretic Methods

Several methods based on information theory have been used for reverse engineering cellular networks from microarray expression profiles. These methods calculate mutual information (MI) between pairs of gene expression profiles. An advantage of MI over other measures of relatedness is that it can detect non-linear interactions. Although these algorithms can be used on time-series data, the sequential aspect is lost, as each sample time point would be considered a different condition.

The original algorithm, relevance networks (RELNET) (Butte and Kohane 2000), infers an interaction if MI for a pair is larger than a threshold. RELNET has been applied to reconstruct networks in yeast (Butte and Kohane 2000), in cancer cell lines (Butte et al. 2000), in human hepatoma cells (Moriyama et al. 2003), and to identify hub cancer genes (Jiang et al. 2008). This approach, however, can result in many false positives, thus extensions which discriminate between direct and indirect interactions have been developed.

Extensions to RELNET proceed in two steps. The first is common to all methods, and consists of calculating MI between pairs of gene expression profiles. In the second step the MI values are assessed and compared, and interactions inferred. The second step is unique to each method.

Context Likelihood of Relatedness (CLR) (Faith et al. 2007) is an algorithm that removes false correlations by comparing MI for each pair with a background distribution of MI scores. CLR was used to reconstruct parts of the transcriptional regulatory network of the pathogen *Salmonella typhimurium* (Taylor et al. 2009).

A second algorithm based on relevance networks, the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) (Basso et al. 2005; Margolin et al. 2006a; Margolin et al. 2006b), eliminates indirect relationships by using data process inequality (DPI), a characteristic of mutual information. ARACNE has been used in reverse engineering the regulatory networks of human B cells (Basso et al. 2005), in the identification of the targets of the transcriptional repressor BCL6 (Basso et al.), in the reconstruction of red blood cell metabolism from metabolic data (Nemenman et al. 2007), and in the genome-wide reconstruction of the regulatory networks of *Streptomyces coelicolor* (Castro-Melchor et al. 2010), an antibiotic producer. CLR and ARACNE were both used to identify genes regulated by Nrf2 in response to oxidative stress (Taylor et al. 2008), and to infer the connectivity of phosphorylation sites in receptor tyrosine kinases (Ciaccio et al.).

A third algorithm, Minimum Redundancy Networks (MRNET) (Meyer et al. 2007), performs a series of maximum relevance/minimum redundancy (MRMR)

selection procedures for each gene and selects the genes having the highest MI with the target.

RELNET, CLR, ARACNE, and MRMR are included in the R package minet (Mutual Information NETwork inference) (Meyer et al. 2008). The networks resulting from these algorithms can be visualized using the R package Rgraphviz (Carey et al. 2005). In addition, the Java implementation of ARACNE includes Cytoscape (Shannon et al. 2003) for network visualization.

3.10.1.2 Bayesian Networks

Bayesian networks have recently emerged as promising approaches for inferring gene regulatory networks using microarray data. These methods are particularly suitable for the reconstruction of cellular networks due to their capability to capture the stochastic nature of gene regulation and allow causality inference (Kim et al. 2003; Murphy and Mian 1999). Furthermore, prior knowledge can be incorporated to improve the accuracy of the final network structure.

A Bayesian network can be represented as a directed acyclic graph, in which each node is a gene, and the edge between two nodes denotes the dependency between two corresponding genes (Heckerman 1998; Needham et al. 2006). A joint probability for the network is thus calculated as a product of multiple conditional probabilities for each gene, given that it is regulated by a defined set of parent genes. These probability functions can be either discrete, e.g., binomial distributions, or continuous, e.g., normal density function. Among the several possible networks being reconstructed, an optimal network can be chosen by maximizing the corresponding posterior probability.

Bayesian predictive networks were constructed using gene expression in combination with genotypic, transcription factor binding site, and protein-protein interaction data in yeast (Zhu et al. 2008). These networks were shown to successfully predict regulators causing hot spots of gene expression activity in a dividing yeast population. Molecular mechanisms underlying transcriptome reprogramming in cyanobacteria under altered environments were also revealed using Bayesian networks (Singh et al. 2010). A large number of genes in the core transcriptional response (CTR)

are associated with oxidative stress under most perturbations, indicating the important role of reactive oxygen species in the regulation of these genes.

3.10.2 Network Reconstruction from Dynamic Gene Expression Data

3.10.2.1 Information Theoretic Method: TimeDelay-ARACNE

Some of the algorithms originally used for network reconstruction using static sampling data have been extended to take advantage of the dependency information contained in time-series data. One such example is an extension of ARACNE. This extension, implemented in the TimeDelay-ARACNE algorithm (Zoppoli et al.), uses time-course data to retrieve time statistical dependencies between gene expression profiles. This algorithm considers the possibility that the expression of a gene at a certain time could depend on the expression level of another gene at an earlier time point, that is, it detects time-delayed dependencies. The algorithm performs three steps: first it detects the time point of the initial changes in the expression for all genes, second it constructs networks by calculating time-dependent MIs, and third it performs network pruning using DPI. TimeDelay-ARACNE, which has been implemented in R, also attempts to infer edge directionality.

3.10.2.2 Dynamic Bayesian Networks (DBNs)

Built upon Bayesian networks, DBNs also calculate a joint probability using the conditional probability of each gene, and select the optimal network based on the posterior probability. DBNs further allow time delay and modeling of feedback loops by incorporating temporal information associated with time-series data. For instance, the cyclic regulation among genes g_a, g_c, and g_d shown in Figure 13a can be represented by allowing these genes to cross-interact from time point i to time point (i+1) (Figure 13b). To further enhance the prediction accuracy and reduce the computational complexity of DBNs, a number of modifications have also been proposed (Zou and Conzen 2005). For example, potential regulators are limited to those genes with either preceding or simultaneous expression changes. Transcriptional time lags between regulators and target genes can also be estimated, and statistical analysis is thereby restricted within that time frame to improve the accuracy of the prediction.

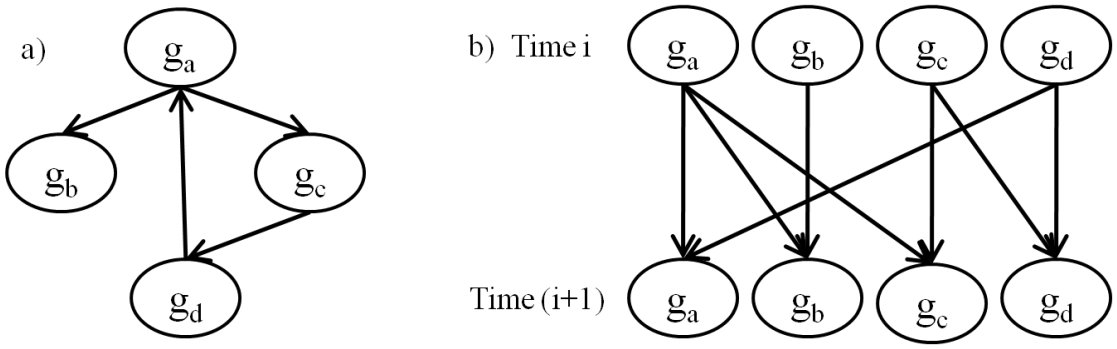


Figure 13: Gene regulatory network with feedback loop deciphered using DBN.

a) A regulatory network containing four genes (g_a, g_b, g_c, and g_d), three of which form a feedback loop ($g_a \rightarrow g_c \rightarrow g_d \rightarrow g_a$).

b) The feedback loop among g_a, g_c, and g_d shown in the left panel is deciphered by allowing cross-interactions along the time axis. For instance, the expression level of g_c at time point (i+1) is dependent on that of g_a at time point i. Similarly, the expression level of g_d at time point (i+1) is dependent on that of g_c at time point i. The loop is closed by allowing g_d's expression level at time point i to have an effect on that of g_a at time point (i+1). Note that each gene's expression level at a certain time point is always dependent on its own expression level at the previous time point.

DBNs have been used successfully to construct gene regulatory networks in yeast using cell cycle time-series microarray data in two independent studies (Kim et al. 2003; Zou and Conzen 2005). Main regulatory nodes in the S.O.S. DNA repair network in *E.coli* were also extracted using DBNs (Perrin et al. 2003). Compared to other methods for inferring gene regulatory networks such as Granger causality and probabilistic Boolean network, DBNs consistently displayed enhanced performance. This was especially the case for short time series, as exemplified with data obtained from muscle development in fruit fly (Li et al. 2007a), normal and infected *Arabidopsis* leaves (Zou and Feng 2009), and food intake effect on human blood (Zhu et al. 2010). Furthermore, the causality inference power of DBNs was substantially improved when time-series gene perturbation data was also incorporated (Dojer et al. 2006).

3.11 CONCLUDING REMARKS

In this chapter, methods for the analysis of microarray data were summarized, with a focus on their use in mammalian cell culture. Whereas specific algorithms used for each step depend on the type of data and the question being asked, the general steps for microarray data analysis remain constant. These steps include data pre-processing followed by identification of differentially expressed genes at a minimum, but greater

biological insight can be gained by using other types of analysis such as profile pattern recognition, pathway analysis, and network reconstruction.

Even though the number of transcriptome studies of antibody producing cell lines has been relatively small compared to other cell types, the next few years will see an increase in the resources available for studying genomes and transcriptomes. This expansion will greatly benefit the understanding of these relevant cell lines.

4 ANALYSIS OF LARGE-SCALE MANUFACTURING DATA FOR ENHANCED PROCESS PERFORMANCE

4.1 SUMMARY

Multivariate analysis of cell culture bioprocess data has the potential of unveiling hidden process characteristics and providing new insights into factors affecting process performance. This chapter presents an investigation of time-series data from 134 process parameters acquired throughout the seed train and the production bioreactors of 243 runs at the Genentech's Vacaville manufacturing facility. Two multivariate methods, kernel-based support vector regression (SVR) and partial least square regression (PLSR), were used to predict the final antibody concentration and the final lactate concentration.

Both product titer and the final lactate level were shown to be predicted accurately when data from the early stages of the production scale were employed. Using only process data from the seed train, the prediction accuracy of the final process outcome was lower; the results nevertheless suggested that the history of the culture might exert significant influence on the final process outcome. The parameters contributing most significantly to the prediction accuracy were related to lactate metabolism and cell viability in both the production scale and the seed train. Lactate consumption, which occurred rather independently of the residual glucose and lactate concentrations, was shown to be a prominent factor in determining the final outcome of production-scale cultures.

The results suggest possible opportunities to intervene in metabolism, steering it towards the type with a strong propensity towards high productivity. Such intervention could occur in the seed stage or in the early stage of the production-scale reactors. Overall, this study presents pattern recognition as an important process analytical technology (PAT). Furthermore, the high correlation between lactate consumption and high productivity can provide a guide to apply quality by design (QbD) principles to enhance process robustness.

4.2 INTRODUCTION

In recent years, cell culture bioprocessing has seen a tremendous growth in data generation and collection. In modern manufacturing facilities, it is not uncommon to encounter hundreds of process parameters being monitored and acquired automatically every few seconds throughout the entire production train. This enormous volume of data further accumulates across multiple campaigns and at multiple manufacturing sites. Mining these historical data holds promise to gain insights into fluctuations in process performance, uncover hidden characteristics of high-performing cultures, and discern process parameters with pivotal contributions to the overall process performance.

Cell culture bioprocess data, however, pose significant challenges to mining practices due to the inherent heterogeneities in time scale and data type (Charaniya et al. 2008). Yet many have successfully applied an array of classification and prediction techniques to investigate hidden process patterns. Principal component analysis (PCA), partial least square regression (PLSR), and other unsupervised techniques, which have the advantage of capturing the interactions among process parameters, have been used for detecting state transitions related to product and lactate formation, online monitoring, fault detection and diagnosis, scale-up assessment, process characterization, and root cause analysis (Bachinger et al. 2000; Gunther et al. 2007; Kirdar et al. 2008; Ündey 2004). In other studies, powerful supervised approaches such as decision tree (DT), artificial neural network (ANN), and support vector regression (SVR) were used to optimize a control scheme incorporating time-course data, predict the final process outcome, and reveal key parameters (Buck et al. 2002; Charaniya et al. 2010; Coleman and Block 2006). Among these multivariate analysis approaches, PLSR and SVR appear to be well-suited to handle the various challenges associated with bioprocess data, namely high-dimensionality and co-linearity between various parameters.

Among the important contributors to differentiating between high- and low-productivity runs of a cell culture process are parameters related to lactate metabolism, including pH, base addition, osmolarity, dissolved CO_2, and lactate concentration (Charaniya et al. 2010). Excessive lactate accumulation has long been known to be an impediment to achieving high cell concentration and superior productivity (Glacken et al.

1986; Hu et al. 1987). Introducing metabolic shifts (i.e., controlling lactate production at low levels or, to a further extent, inducing lactate consumption) has been performed using various strategies. These approaches include dynamic feeding to control glucose at low levels (Cruz et al. 1999; Zhou et al. 1997), using alternative carbon sources (Altamirano et al. 2006; Wlaschin and Hu 2007a), knocking down LDH-A (Chen et al. 2001; Kim and Lee 2007a), and enhancing glucose carbon flux into the TCA cycle (Irani et al. 1999; Kim and Lee 2007b). Understanding the linkage between lactate metabolism and high productivity thus offers the opportunity to discover the metabolic signatures of these high-performing processes.

In this study, we employed support vector regression (SVR) and partial least square regression (PLSR) methods to predict the final process outcome using process data from 243 production runs at a Genentech manufacturing facility. This dataset comprises 134 temporal parameters acquired online and offline throughout the seed train (80 L, 400 L, and 2000 L) and the production-scale bioreactors (12000 L). Parameters pivotal to prediction accuracy were assessed based on two criteria: the frequency of occurrence (f) in the best parameter sets for SVR models and the magnitude of the regression coefficient (β) in the optimal PLSR models. Among these pivotal parameters, various aspects of the lactate consumption phenomenon at the production scale in high-titer runs were further investigated.

4.3 MATERIALS AND METHODS

4.3.1 Overview of Mammalian Cell Culture Processes

The large-scale manufacturing bioreactors from which bioprocess data were acquired were located at the Genentech's Vacaville facility. The recombinant hIgG-producing CHO cells were thawed from a banked vial and expanded in shake flasks of increasing sizes before reaching 20 L bioreactors (Figure 14). This rolling seed was maintained at 20 L scale to inoculate multiple seed trains, each of which comprises three scales: 80 L, 400 L, and 2000 L. The cells were expanded step-wise throughout the seed bioreactors, each of which lasted for approximately 72 hr in a batch mode. Subsequently, the cells were transferred to the production-scale bioreactors of 12000 L, in which they

were cultured for about 264 hr in a fed-batch mode with feeding and temperature shift at 72 hr.

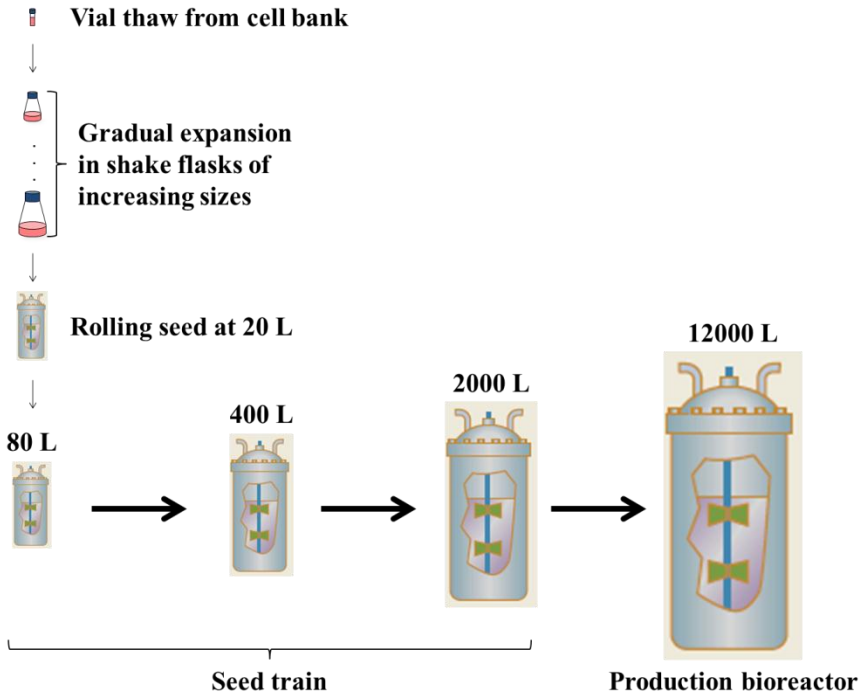


Figure 14: Overview of a production train at the Genentech's Vacaville manufacturing facility.

After thawing and expansion in shake flasks, recombinant cells were maintained at 20 L to inoculate multiple production trains, each comprising a seed train (80 L, 400 L, and 2000 L) and a production (12000 L) bioreactor. At each of the seed bioreactors, the cells were expanded in a batch mode for 72 hr. At the production bioreactor, they were cultured in a fed-batch mode for 264 hr with feeding and temperature shift at 72 hr.

4.3.2 Collection of Large-scale Manufacturing Data

For each runs, hundreds of process parameters were measured at various frequencies over the course of the production train. The final hIgG antibody concentration, also known as final titer, was quantified at the end of the production stage and normalized to an average value of 1.00. Hundreds of online parameters were recorded electronically every minute, whereas tens of offline parameters were determined by periodic withdrawal of samples from the bioreactors approximately every 24 hr. In addition, several specific rates were calculated from these measured parameters. A total number of 134 process parameters were selected for analysis as listed in Table 1.

Table 1: List of 134 temporal process parameters used in the analysis.

Thirty-three parameters were collected at each seed scale (80 L, 400 L, and 2000 L), and thirty-five parameters at the production scale (12000 L).

Offline parameters	**Online parameters**
Ammonium ion concentration	Air sparge rate
Dissolved CO_2 (pCO_2)	Air sparge set point
Dissolved O_2 (pO_2)	Backpressure (12000 L only)
Glucose concentration	CO_2 sparge rate
Integrated packed cell volume (IntvPCV) (12000 L only)	Dissolved oxygen (DO) controller output
Lactate concentration	DO (primary)
Osmolarity	DO (secondary)
Packed cell volume (PCV)	Flowrate overlay
pH (offline)	Jacket temperature
Sodium ion concentration	O_2 sparge rate
Viability	pH controller output
Viable cell density (VCD)	pH (online)
Derived parameters	Pressure exhaust valve
Specific cell growth rate (μ)	Reactor weight
Specific glucose consumption rate (q_{Lac})	Total air sparged
Specific lactate consumption rate (q_{Glc})	Total base added
	Total CO_2 sparged
	Total O_2 sparged
	Total gas sparged
	Vessel temperature

4.3.3 Data Pre-processing

Online data acquired at each scale were smoothed using a moving window average method with a time window of 100 min. At every time point, a parameter's value was approximated as the average of all measurements of that parameter within the time window. For instance, the pre-processed value at time *t* is the average of measurements at times *t, t+1, t+2, …, t+99.* The raw and the pre-processed temporal profiles of CO_2 sparge rate and DO controller output of one run at 12000 L scale are shown in Figure 15a as examples.

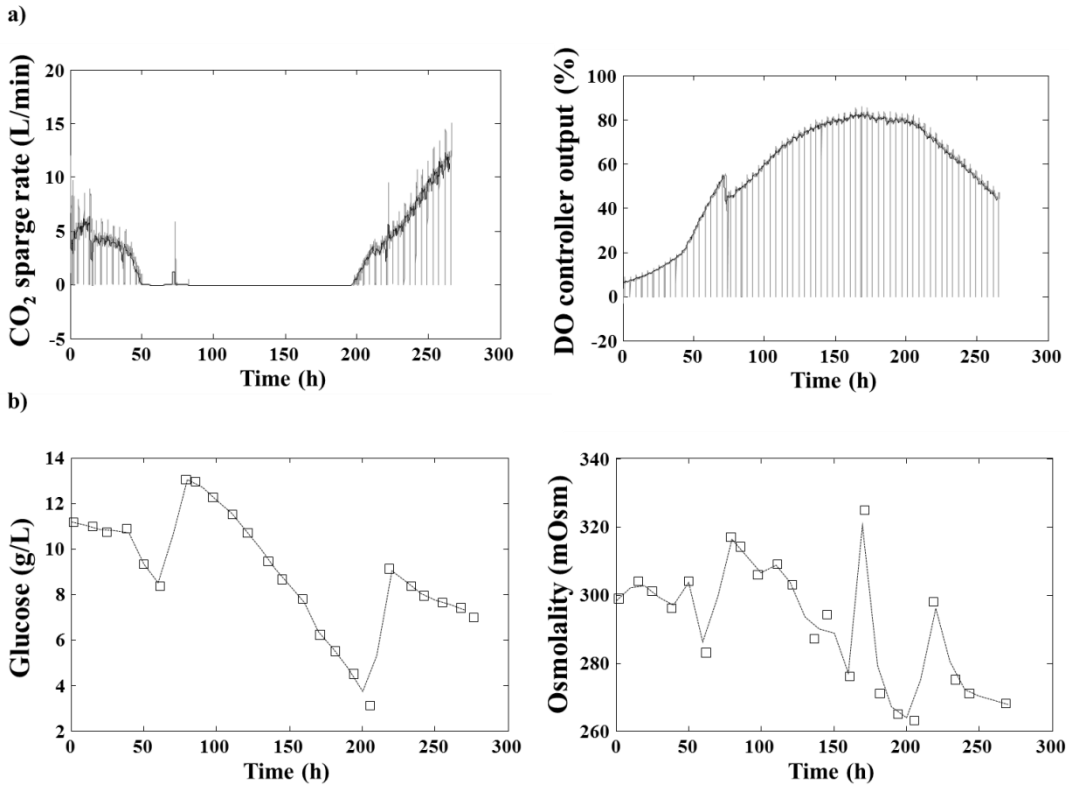


Figure 15: Pre-processing of cell culture bioprocess data.

(a) Online parameters were pre-processed using a moving window average method: (——) measured, (——) pre-processed.
(b) Offline parameters were pre-processed using a linear interpolation scheme: (□) measured, (——) interpolated.

Due to differences in sampling frequency, offline parameters were linearly interpolated and extrapolated every 20 hr. Figure 15b shows the temporal profiles of glucose concentration and osmolality in the production bioreactor for one example run. Furthermore, calculated specific rates of lactate production, glucose consumption, and cell growth were smoothed using third-order polynomials.

4.3.4 Stage-wise Organization of Data

Process data from all scales were organized into eight individual and seven cumulative datasets as shown in Figure 16 to investigate progression of the production trains. The first individual dataset comprised process data from the 80 L scale bioreactors. The second dataset contained data from the next scale of 400 L, and so on.

Since the run time at the production scale (260 hr) was much longer compared to that at each of the seed scales (70 hr), it was segregated into several stages: up to 70 hr, 120 hr, 170 hr, 220 hr, and 260 hr. In addition to these eight individual datasets, process data were also accumulated across scales with the largest dataset compiling data from 80 L, 400 L, 2000 L, and up to 260 hr of the 12000 L scale.

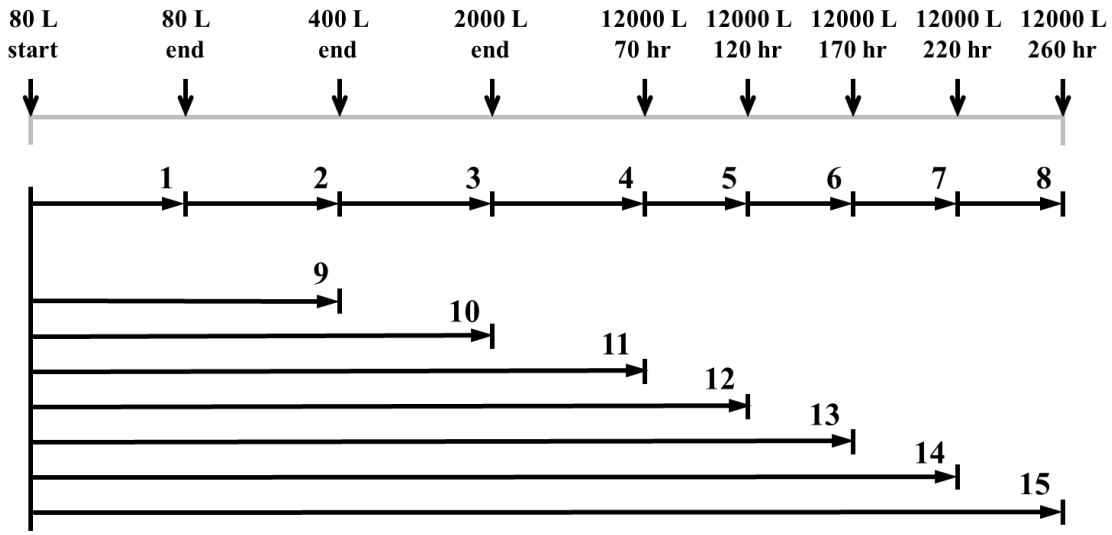


Figure 16: Stage-wise organization of process data into fifteen datasets.

Eight individual datasets (from 1 to 8) comprised data from each seed bioreactor or each stage (up to 70 hr, 120 hr, 170 hr, 220 hr, and 260 hr) of the production bioreactor. Seven other datasets (from 9 to 15) incrementally accumulated data across time.

4.3.5 Model Training and Evaluation Using 10-Fold Cross-Validation

A 10-fold cross-validation scheme as shown in Figure 17 was used for training and evaluation of both support vector regression (SVR) and partial least square regression (PLSR) models. Process data from 243 runs in each of the fifteen datasets described above were randomly divided into ten subsets of approximately equal sizes. During each round of cross-validation, nine of the ten subsets were used as the training set on which model optimization was performed. The best performing model on each training set was used to predict process outcome of runs in the corresponding, unseen test set (10^{th} subset). This process was repeated ten times on different pairs of training and test subsets.

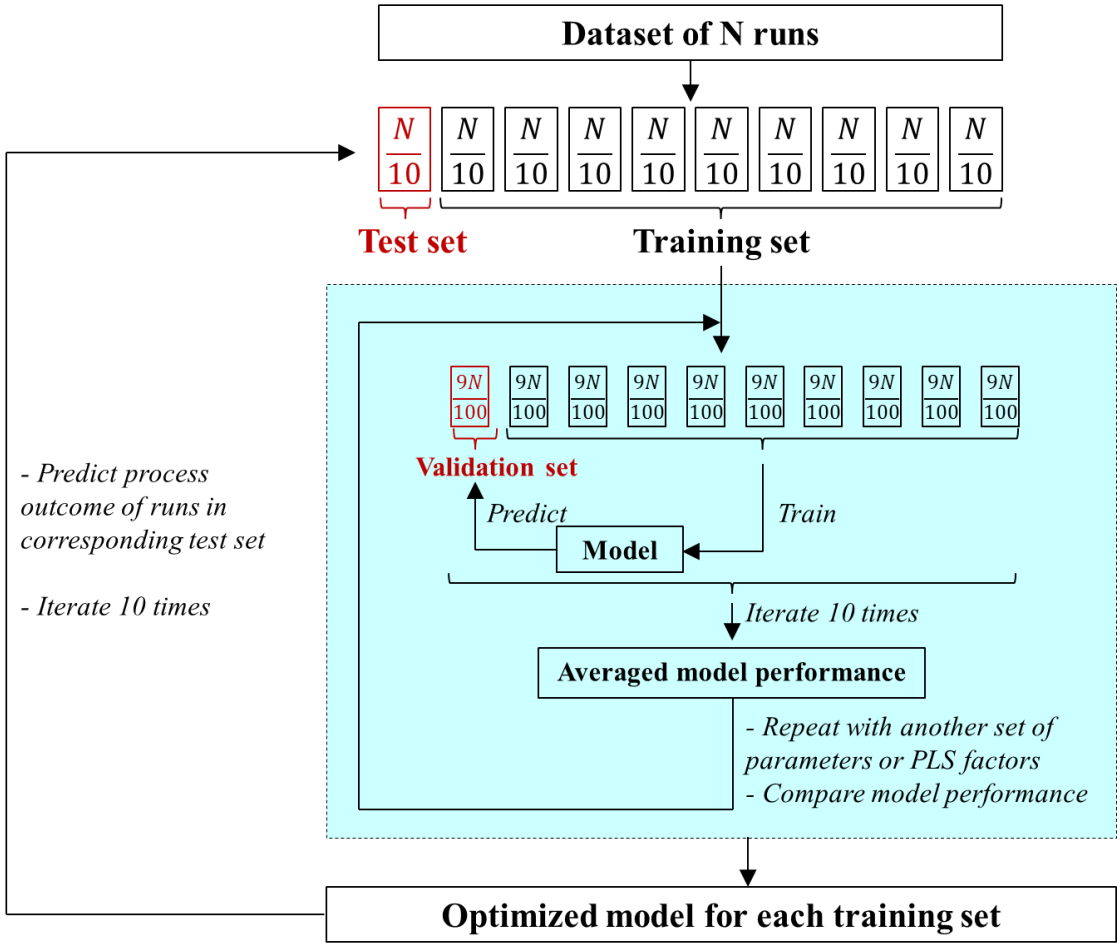


Figure 17: Ten-fold cross-validation scheme with model optimization for both multivariate approaches.

All n process runs ($n = 243$ in this study) were randomly separated into ten equal subsets. Nine were used as the training set on which model optimization was performed. The optimized model was used to predict the outcome of runs in the 10^{th} subset (test set). Model optimization involved further random separation of the training set into 10 equal groups. Again, nine were used to train a model with a certain set of parameters (for SVR approach) or PLS factors (for PLSR approach). The performance of this model was tested on the 10^{th} group (validation set). This process was repeated 10 times to obtain the average performance for each set of parameters/factors, which was later compared to identify the parameter/factor set that resulted in the best predictive (optimized) model. The shaded box contains all steps in model optimization.

Model performance was evaluated using the Pearson's correlation coefficient (r) and the root mean square error (ε) between the predicted and the actual final process outcome:

Equation 1: Pearson's correlation coefficient r between predicted ($f(x_i)$) and actual outcome (y_i).

$$r = \frac{\sum_{i=1}^{n} y_i f(x_i) - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} f(x_i)}{n}}{\sqrt{\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} f(x_i)^2 - \frac{\left(\sum_{i=1}^{n} f(x_i)\right)^2}{n}\right)}}$$

Equation 2: Root mean square error ε between predicted ($f(x_i)$) and actual outcome (y_i).

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} (y_i - f(x_i))^2}{n}}$$

where n is the number of runs, and y_i and $f(x_i)$ are the actual and the predicted titer values of run i, respectively. The model performance was averaged across the 10 folds. As a baseline for evaluating model performance, a random predictor with one million simulations of randomized final process outcome was generated.

To get a better estimate of the generalization error of the constructed models, model optimization (i.e. selection of model and process parameters) was further performed on each training set, also using 10-fold cross-validation, as shown in the shaded box in Figure 17. Model optimization was performed for each round of the 10-fold cross-validation. This involved further partitioning of the training set randomly into ten smaller groups of about equal sizes. The model was trained on nine groups using a different set of parameters for SVR approach or PLS factors for PLSR approach. The performance of the resulting model was subsequently tested in the 10^{th} group, called the validation set. This procedure was repeated ten times for each set of parameters or PLS factors. The average performance of the model over these inner 10 folds was used to determine the optimal set of parameters or PLS factors for each round of the outer 10-fold cross-validation. Subsequently, the best model was selected and used to predict the outcome of runs in the corresponding, unseen test set.

4.3.6 Construction of Partial Least Square Regression (PLSR) Models

Partial least square regression (PLSR) models were constructed using the SIMPLS algorithm (Chong and Jun 2005; de Jong 1993). Time-series data for each process parameter were extracted every 10 hr, resulting in multiple discrete "variables" originating from the same parameter. These variables were concatenated over the run

time of each scale into a data matrix ($\boldsymbol{X}$). Data in each column of this matrix were further autoscaled to a mean of zero and a standard deviation of one to give a new matrix $\boldsymbol{X0}$. A similar transformation was also performed on the response vector ($\boldsymbol{y}$) to obtain the autoscaled final process outcome ($\boldsymbol{y0}$) (either antibody titer or lactate concentration at the end of the 12000 L cultures).

The autoscaled data matrix ($\boldsymbol{X0}$) was projected onto mutually orthogonal PLS factors ($\boldsymbol{XS}$), each of which is a weighted linear combination of the original variables in $\boldsymbol{X0}$. A set of these PLS factors can be used to construct a regression function to predict the autoscaled final process outcome in $\boldsymbol{y0}$. The SIMPLS algorithm for a univariate response in $\boldsymbol{y0}$ can be simplified in the following equations:

Equation 3: Projection of original data *X*0 into mutually orthogonal PLS factors *XS*.

$$\boldsymbol{XS}_{n\times a} = \boldsymbol{X0}_{n\times p}\,\boldsymbol{W}_{p\times a}$$

$$\boldsymbol{X0}_{n\times p} = \boldsymbol{XS}_{n\times a}\,\boldsymbol{XL}_{p\times a}{}^{T} + \boldsymbol{XE}_{n\times p}$$

Equation 4: Prediction of process outcome *y*0 using PLS factors in *XS*.

$$\boldsymbol{y0}_{n\times 1} = \boldsymbol{XS}_{n\times a}\,.\boldsymbol{\beta}_{a\times 1} + \boldsymbol{ye}_{n\times 1}$$

such that the covariance between $\boldsymbol{X0}$ and $\boldsymbol{y0}$ is maximized.

In these equations, $\boldsymbol{XS}$, $\boldsymbol{X0}$, and $\boldsymbol{W}$ are the matrix of orthogonal PLS factors, the autoscaled data matrix, and the matrix of PLS weights, respectively. The matrices $\boldsymbol{XL}$ and $\boldsymbol{XE}$ contain loadings of the PLS factors and the residuals when factorizing $\boldsymbol{X0}$ into a product of $\boldsymbol{XS}$ and $\boldsymbol{XL}^{\boldsymbol{T}}$, respectively. The vectors $\boldsymbol{y0}$, $\boldsymbol{\beta}$, and $\boldsymbol{ye}$ comprise the autoscaled response, the regression coefficients of $\boldsymbol{y0}$ using $\boldsymbol{XS}$, and the residuals when regressing $\boldsymbol{y0}$ using $\boldsymbol{XS}$, respectively. The variables $n, p,$ and a are the number of process runs, the number of variables (in this case, a product between the number of process parameters m and the number of time points t), and the number of PLS factors used for regression, respectively.

The *plsregress* subroutine, an implementation of the SIMPLS algorithm in the Matlab's statistics toolbox, was used for constructing the PLSR models. For each of the fifteen datasets, a PLSR model was constructed and optimized as described in Model

Training and Evaluation Using 10-Fold Cross-Validation. The number of PLS factors in each model was varied from one to the maximum possible (which is the rank of the data matrix $\boldsymbol{X0}$). For each fold of the outer 10-fold cross-validation, an optimal set of PLS factors, and thus variables, could be identified. Furthermore, as each original parameter was discretized into multiple variables, the average magnitude of the regression coefficients of all variables which originated from the same parameter was used to assess the importance of that parameter.

4.3.7 Construction of Support Vector Regression (SVR) Models

LIBSVM (Chang and Lin 2001), an implementation of the SVR algorithm in C, was used to construct v-SVR models. This algorithm utilized the similarity measure between runs, which were calculated using the following two-step approach, as the input. In the first step, the Euclidean distance over time between any two runs *i* and *j* was computed for each individual parameter p_m:

Equation 5: Euclidean distance between run i and run j for parameter p_m.

$$d_{ij}^{p_m} = \sqrt{\sum_t (p_{m,it} - p_{m,jt})^2}$$

Subsequently, this distance was scaled to 0 – 1 and converted into a similarity value:

Equation 6: Similarity measure between run i and run j for parameter p_m.

$$\boldsymbol{s_{ij}^{p_m} = 1 - d_{ij}^{p_m}}$$

All pair-wise similarity values between runs were organized into an $n \times n$ matrix, where n is the number of runs.

In the second step, the similarity matrices of all parameters were linearly combined to form a final similarity matrix, which was used as a pre-defined kernel in the v-SVR algorithm. Upon combination, each parameter was either given equal weights of $1/m$ (where m is the number of process parameters) or weighted according to how well it correlates to the final process outcome using the non-linear Spearman's rank correlation coefficient:

Equation 7: Overall similarity measure between run *i* and run *j* taking all parameters into consideration.

$$s_{ij} = \sum_m w_m s_{ij}^{p_m}$$

where $w_m = 1/m$ for equal weighting or $w_m = -corr(s_{ij}^{p_m}, \Delta y_{ij})$ with Δy_{ij} denotes the difference in the final process outcome between run i and run j. Thus all entries in the final similarity matrix were maintained between 0 and 1. The objective function (y), either the final titer or the final lactate concentration, was also scaled to the same range of 0 – 1.

A v-SVR model seeks to identify a regression function $f(\boldsymbol{x}) = \boldsymbol{wx} + \boldsymbol{b}$ that minimizes the prediction error ε (Equation 2). Differences exceeding ε are penalized by a slack variable (ξ_i or ξ_i') and an *a priori* chosen cost function (C). The parameters ($\boldsymbol{w}, \boldsymbol{b}$) of the regression function are obtained by solving the following constrained optimization problem:

Equation 8: Optimization problem formulated in v-SVR.

$$\min_{\boldsymbol{w},\boldsymbol{b}} \left\{ \frac{1}{2} \|\boldsymbol{w}\|^2 + C \left(v\varepsilon + \frac{1}{n} \sum_{i=1}^{n} (\xi_i + \xi_i') \right) \right\}$$

subject to the following inequality constraints:

Equation 9: Inequality constraints of v-SVR optimization problem.

$$(wx_i + b) - y_i \leq \varepsilon + \xi_i, \forall i$$

$$y_i - (wx_i + b) \leq \varepsilon + \xi_i', \forall i$$

where ξ_i, ξ_i', and $\varepsilon \geq 0$.

The parameter v is a non-negative constant that determines the balance between the complexity of the model and the extent of the prediction error ε. The default value of $v = 0.5$ was used. In addition, a simple grid search within the range of 0 – 1 with 0.1 intervals was performed on the cost function. The best value was used in the subsequent steps of model optimization and identification of pivotal process parameters.

v-SVR models were constructed and optimized for each of the eight individual datasets as described in Model Training and Evaluation Using 10-Fold Cross-Validation. For the seven cumulative datasets, due to computational constraints imposed by the large number of parameters, SVR models were built using the best performing sets of parameters obtained for the corresponding individual datasets.

4.3.8 Identification of Pivotal Process Parameters Using SVR Approach

A greedy parameter selection approach based on the wrapper feature-selection method (Liu and Hiroshi 1998) was used to find the best performing set of parameters for the SVR models. This approach determined the suitability of a set of features (i.e., process parameters) by first building an SVR model using these features and then assessing its performance on a subset of the data that was not used for training (i.e., validation set). The set of features whose model achieved the best performance on the validation set became the set of selected parameters. Since each of the eight individual datasets contains either 33 parameters at the seed scales or 35 parameters at the production scale, a direct application of the wrapper feature-selection method will require an evaluation of $2^{33} - 1$ or $2^{35} - 1$ (excluding the *null* set) possible parameter subsets, which is prohibitively large. For this reason, we employed a greedy strategy that only considers a substantially smaller number of parameter subsets.

In this approach, the different parameter subsets were organized into a lattice structure, whose i^{th} level contained all the subsets of size $(m - i)$ where m is the number of parameters. All nodes at each level were connected to the nodes of the preceding level that were its supersets. The algorithm started by evaluating the performance of the subsets at levels 0 and 1 (i.e., the entire set of m parameters and the m subsets that were obtained by removing one parameter, respectively). Among the m subsets at level 1, N subsets whose models achieved the best performance on the validation set were retained. The algorithm then proceeded to evaluate the performance of subsets at level 2 that are descendants of at least one of the N nodes retained at level 1. Among those subsets, it also retained the N best performing ones. This process continued until the last level of the lattice. Note that, by setting N to a small value (in our

experiments, $N = \{5, 15, 25, 35\}$) and by considering only subsets whose supersets were among the N best performing subsets of the previous level, the total number of subsets being considered became computationally feasible. In addition, since the subsets that were pruned are those that did not perform well, this approach could still identify good performing parameter subsets.

Since a 10-fold cross-validation scheme was performed, each fold generated an optimal set of parameters. Thus the occurrence frequency (f) of each parameter over all 10 folds can be used as an estimate of its contribution to the overall model performance.

4.4 RESULTS

4.4.1 High- and Low-Performing Runs Exhibit Distinct Process Characteristics

The 243 production runs investigated in this study exhibited considerable variation in a number of process parameters and outcome as shown in Figure 18. The pre-harvest recombinant antibody concentration, also known as the final titer, varied across a wide range from 0.70 to 1.25 (Figure 18a). These runs were categorized into three classes: top 20% (in blue), middle 60% (in grey), and bottom 20% (in red), with their final titer approximately over 1.10, between 1.10 and 0.90, and below 0.90, respectively. Because of measurement error of recombinant antibody concentration, it is possible that runs within the middle 60% class have a high degree of similarity. In contrast, comparison of the top 20% and the bottom 20% runs should bear distinct characteristics of high-titer cultures.

As shown in Figure 18b, both the top and bottom 20% cultures started with a similar range of cell concentration in the production-scale bioreactors. There was a substantial spread of cell concentration at peak growth and at the end of the culture, even among runs within the top or bottom 20% class. In general, the top 20% runs reached higher peak cell concentrations (between 100 and 150 hr) although the range was rather wide. It is apparent that more runs of the bottom 20% class had lower peak cell concentrations, and all bottom 20% runs had lower viable cell concentrations at the end of the production run.

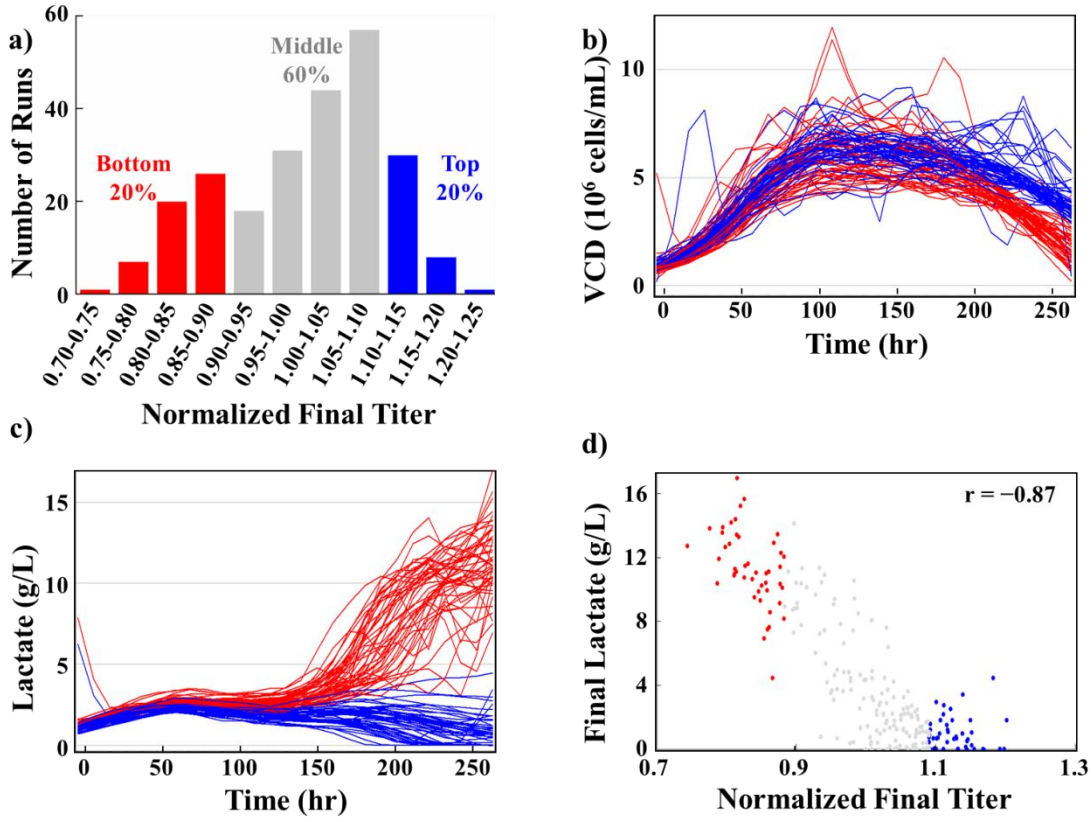


Figure 18: Differences in process performance as indicated by the final antibody concentration (titer), viable cell density (VCD), and lactate concentration across 243 production runs.

- **a) Distribution of the final titer (normalized such that the average across all runs is 1.00). Roughly 20% of runs have final titers greater than 1.10 (top 20% - in blue); 20% of runs have titers less than 0.90 (bottom 20% - in red); and 60% of runs have titers between 0.90 and 1.10 (middle 60% - in grey)**
- **b) Variation in viable cell density at 12000 L scale between runs in the top 20% (blue) and the bottom 20% (red)**
- **c) Variation in lactate concentration at 12000 L scale between runs in the top 20% (blue) and the bottom 20% (red)**
- **d) High correlation (-0.87) between the final lactate concentration and the final titer across all runs**

The lactate concentration profiles also showed profound differences between the top 20% and bottom 20% runs (Figure 18c). Although lactate concentrations were in similar ranges in all cultures initially, by the time cell concentration reached the peak, they had become higher in the bottom 20% runs. Despite a period between 100 and 130 hr during which lactate production subsided, all bottom 20% runs proceeded to return to the lactate production state whereas nearly all top 20% runs switched to the lactate consumption state. Many of these top 20% runs resulted in complete exhaustion of lactate previously produced during the exponential growth stage. As expected, the final

lactate concentration was found to be highly correlated to the product yield in all runs with a Pearson's correlation coefficient of −0.87 (Figure 18d), indicating a close connection between cellular metabolic activities and product titer.

4.4.2 Process Outcome is Predicted Accurately Using Multivariate Models

Two multivariate regression approaches, support vector regression (SVR) and partial least square regression (PLSR), were employed. Time-series process data were acquired for 134 online, offline, and derived parameters throughout the seed train (80 L, 400 L, and 2000 L) and the production-scale bioreactors (12000 L). To investigate the importance of these parameters at each scale, process data were organized into fifteen different datasets as shown in Figure 19. In addition to the final titer, the final lactate concentration was also used as an objective function because of the indication of its being an attribute of a culture's performance. The prediction accuracy of these models was assessed based on the Pearson's correlation coefficient (r) and the root mean square error (ε) as described in Model Training and Evaluation Using 10-Fold Cross-Validation.

In constructing SVR models, a grid search of the cost function in the range of 0 – 1 with 0.1 intervals yielded an optimal value of 1, which was used for constructing subsequent SVR models. Both differential and equal weighting schemes as described in Construction of Support Vector Regression (SVR) Models were employed to combine all similarity matrices. Since the equal weighting scheme resulted in slightly better model performance (data not shown), it was used for the subsequent step of feature selection. A *wrapper-based feature selection* algorithm as described in Identification of Pivotal Process Parameters Using SVR Approach in Methods was further employed to identify the optimal combination of parameters that result in the lowest root mean square error (ε). The top 35 nodes (i.e., $N = 35$) were expanded at each level. Three additional values of $N = \{25, 15, 5\}$ were also evaluated, and resulted in similar performances for the 8th dataset (12000 L up to 260 hr). Thus, for the other individual datasets, the maximum number of nodes to be expanded at each level was fixed at 35.

Similarly, in constructing PLSR models, a 10-fold cross-validation scheme (Figure 17) was used to find the optimal number of PLS factors to be incorporated in

each model that gives the best predicted final process outcome. As described in Model Training and Evaluation Using 10-Fold Cross-Validation, the number of PLS factors was varied from one to the rank of the data matrix $\boldsymbol{X0}$. The optimized number of PLS factors in each training set was used to construct a PLSR model for the corresponding test set.

It is interesting that prediction trends across different datasets were considerably similar irrespective of the multivariate approach as shown in Table 2. Overall, the PLSR approach appeared to result in slightly better models than those constructed using the SVR approach. However, when the input data were noisy (400 L and 2000 L), these PLSR models failed to maintain good performances whereas SVR models built using the same datasets were still robust. Furthermore, similar correlations between the predicted and the actual final process outcome were observed across all datasets regardless of whether the final titer or the final lactate concentration was used as the objective function. This result indicates that product yield and cellular metabolic activities are indeed closely interconnected, confirming the high correlation between these two characteristics as previously shown in Figure 18d. Due to this considerable similarity in prediction accuracy, results are presented for SVR models predicting the final titer herein. It is noteworthy that a random predictor generates a root mean square error of 0.17 and 7.80 for the final titer and the final lactate concentration, respectively, and a Pearson's correlation coefficient of zero in both cases.

Table 2: Prediction accuracy of PLSR and SVR models using process data acquired at different stages.

Model performance was evaluated using the Pearson's correlation coefficient (r) and the root mean square error (ε) between the predicted and the actual final process outcome.

Dataset	Final antibody concentration (titer)				Final lactate concentration			
	PLSR		SVR		PLSR		SVR	
	r	ε	r	ε	r	ε	r	ε
80 L	0.42	0.10	0.40	0.10	0.44	4.18	0.43	4.15
400 L	0.21	0.12	0.43	0.10	0.33	4.79	0.43	4.15
2000 L	0.28	0.11	0.35	0.10	0.28	4.63	0.37	4.30
12000 L up to 70 hr	0.73	0.07	0.73	0.07	0.72	3.18	0.77	2.93
12000 L up to 120 hr	0.80	0.06	0.77	0.07	0.85	2.41	0.78	2.84
12000 L up to 170 hr	0.88	0.05	0.86	0.06	0.95	1.47	0.92	1.98
12000 L up to 220 hr	0.92	0.04	0.91	0.04	0.97	1.09	0.96	1.56
12000 L up to 260 hr	0.92	0.04	0.92	0.04	0.97	1.13	0.98	1.33
80 L+400 L	0.41	0.10	0.47	0.09	0.50	3.97	0.48	4.00
80 L+400 L+2000 L	0.45	0.09	0.48	0.09	0.45	4.10	0.50	3.94
80 L+400 L+2000 L+12000 L up to 70 hr	0.71	0.07	0.68	0.08	0.73	3.13	0.68	3.39
80 L+400 L+2000 L+12000 L up to 120 hr	0.77	0.07	0.72	0.08	0.76	2.94	0.73	3.24
80 L+400 L+2000 L+12000 L up to 170 hr	0.88	0.05	0.83	0.06	0.90	1.97	0.87	2.65
80 L+400 L+2000 L+12000 L up to 220 hr	0.91	0.04	0.91	0.05	0.97	1.18	0.95	2.00
80 L+400 L+2000 L+12000 L up to 260 hr	0.92	0.04	0.92	0.05	0.97	1.14	0.96	1.82

Data acquired at the smallest scale of the seed train (80 L) were moderately indicative of the final titer with a correlation coefficient (r) of 0.40 and a root mean square error (ε) of 0.10 (Figure 19a). The SVR model constructed using data from the next scale of 400 L performed slightly better with $r = 0.43$ and $\varepsilon = 0.10$. Data from

2000 L scale bioreactors, surprisingly, were less informative than data from the two smaller scales. The correlation coefficient dropped to 0.35 and the error remained at 0.10 as shown in Figure 19b. This reduced performance appeared to be circumvented by concatenating data across these scales. The SVR model built upon data concatenated from 80 L and 400 L scales exhibited a slight improvement than did those built using data from each individual scale ($r = 0.47, \varepsilon = 0.09$). Similarly, cumulative data across all three scales of the seed train resulted in a correlation coefficient of 0.48 and an error of 0.09.

When data from the first 70 hr of the production scale was used, the prediction accuracy increased sharply to 0.73 with a root mean square error of 0.07 (Figure 19c). By the time most runs reached peak growth at 120 hr, the performance improved to $r = 0.77$ and $\varepsilon = 0.07$. As the runs approached the end, the final titer could be predicted with higher correlation coefficients of 0.86 ($\varepsilon = 0.06$) by 170 hr, and 0.92 (ε = 0.04) upon completion at 260 hr (Figure 19d). Interestingly, the regression models built upon data acquired at the production scale alone were slightly more predictive compared to those with the addition of data from the seed train. Concatenating data from the seed train to the first 70 hr of the production scale actually reduced the prediction accuracy from $r = 0.73$ and $\varepsilon = 0.07$ to $r = 0.68$ and $\varepsilon = 0.08$. At around peak growth (~120 hr), addition of seed data did not result in a better model ($r = 0.72, \varepsilon = 0.08$ compared to $r = 0.77, \varepsilon = 0.07$). Similarly, concatenating seed data with data from the late stage of the production scale also did not improve prediction accuracy. The model built upon concatenating all data showed little to no improvement ($r = 0.92, \varepsilon = 0.05$) over the model built on data from the production scale only ($r = 0.92, \varepsilon = 0.04$). This result suggests that the seed data are rather noisy relative to the production-scale data and incorporation of these data may not help increase model prediction accuracy.

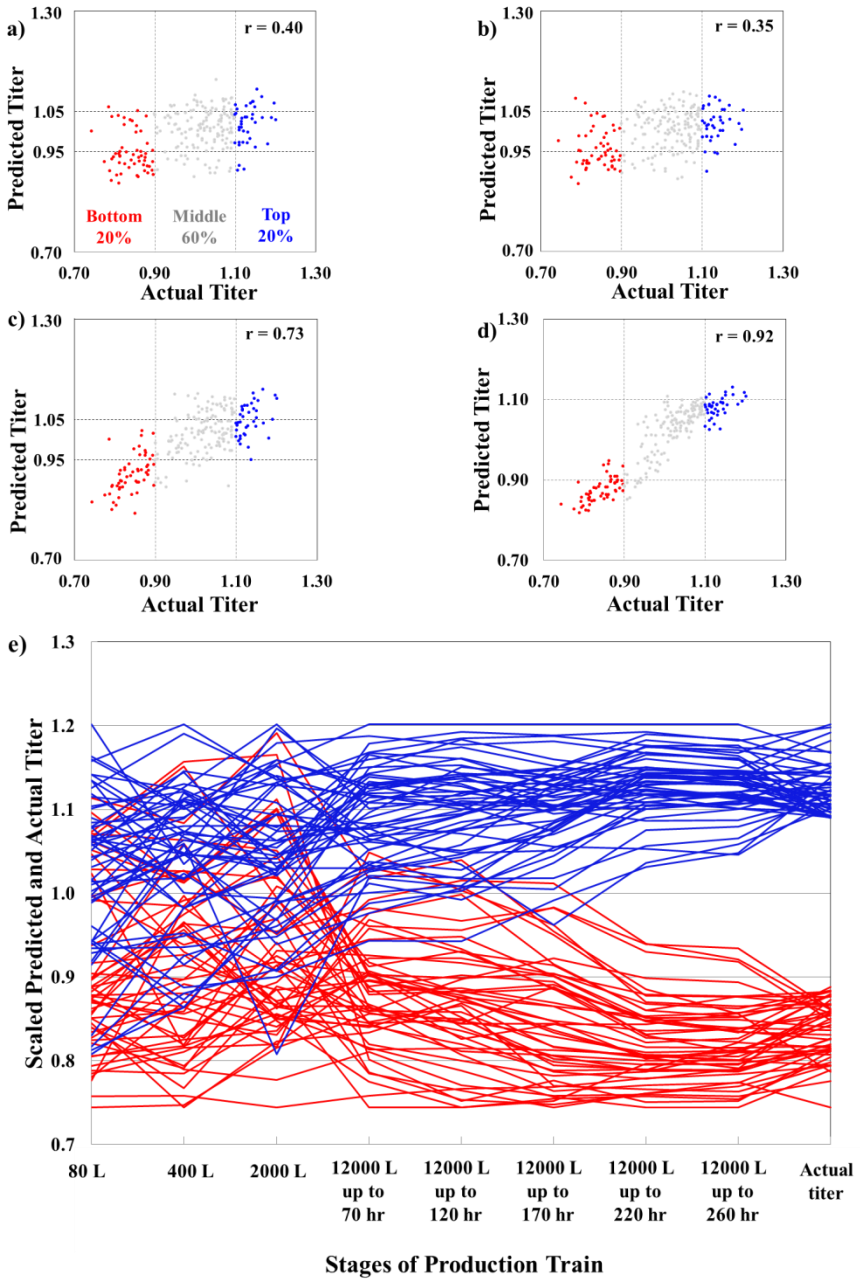


Figure 19: SVR models' prediction accuracy of the final titer using different datasets.

The correlation coefficient (*r*) between the predicted and the actual titer is shown for each dataset. The dashed lines indicate the separation of the top 20%, middle 60%, and bottom 20% of runs based on the predicted titer (y-axis) or the actual titer (x-axis). Runs in the top 20% class are colored in blue; runs in the middle 20% class are colored in grey; and runs in the bottom 20% class are colored in red.

- **a) 80 L scale**
- **b) 2000 L scale**
- **c) Up to 70 hr of 12000 L scale**
- **d) Up to 260 hr of 12000 L scale**
- **e) The progression of predicted titer is shown over the course of the cultures for the top and the bottom 20% runs**

Furthermore, as evident from Figure 19a and b, using data from the 80 L and 2000 L scales, only a few runs predicted to be in the top 20% class (above the horizontal grid line of $y = 1.05$) actually fell to the bottom 20% class of the actual titer (on the left of the vertical grid line of $x = 0.90$). Similarly, the number of runs predicted to be in the bottom 20% class (below the horizontal grid line of $y = 0.95$) that ended up in the top 20% class (on the right of the vertical grid line of $x = 1.10$) was also small. Once data from the production scale, even as early as the first 70 hr was used, this class switch was not observed in any runs (Figure 19c and d). This result indicates that process characteristics at the early stage of the production scale are already indicative of the final outcome, and no runs are inclined to switch between the top and the bottom classes after this stage.

We next examined those few runs which switched classes by tracking their performance over the course of the run. The performance as judged by titer predicted using each of the eight individual datasets is shown in Figure 19e. For better visualization, the titer values predicted using each dataset were linearly scaled such that they were in the same range as the actual titer values throughout. Again red and blue colors indicate the bottom and the top 20% of runs, respectively. It is interesting to note that class switch occurred relatively gradually over different stages of the seed train (80 L, 400 L, and 2000 L). By 70 hr of the production scale, switching has virtually completed. This result clearly points out that any intervention means should be carried out during the seed train or at least by the first 70 hr of the production scale.

4.4.3 Majority of Pivotal Parameters are Related to Cell Growth and Lactate Metabolism

The contribution of each parameter to the prediction of the final process outcome was assessed using two criteria: the magnitude of the regression coefficient (β) in the optimized PLSR models and the frequency of occurrence (f) in the selected parameter sets for SVR models. As described in Materials and Methods, a *wrapper-based feature selection* algorithm was employed to identify the minimum combination of parameters that results in an SVR model with the lowest validation error. Shown in Figure 20 is this

error as a function of the number of parameters incorporated into SVR models at each scale.

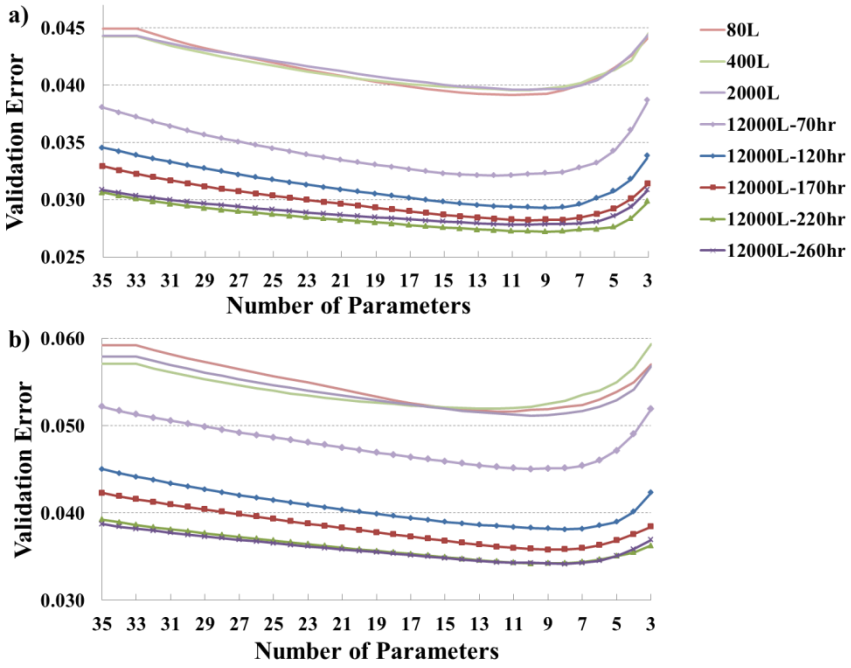


Figure 20: Variation of validation error as a function of the number of parameters used in SVR models.

a) Final titer as the objective function
b) Final lactate concentration as the objective function

Initially, the SVR models appeared to perform better with the gradual removal of parameters, indicating that most of these parameters are indeed redundant or even noisy. In most cases, the best model was constructed using a set of six to eight parameters as indicated by the valley in the validation error profile. The immediate, sharp rise of error following the removal of parameters in this selected set from the model suggests that they play a pivotal role in model prediction accuracy. Thus the occurrence frequency (f) of each parameter in all selected sets represents its relative contribution to the SVR models' performance.

As shown in the previous section of Process Outcome is Predicted Accurately Using Multivariate Models, class switch appeared to occur either during the seed train or by 70 hr of the production scale. We thus focused on identifying pivotal parameters at

these two stages to search for possible hints of intervention. Figure 21 shows the relative importance of 33 process parameters acquired at the smallest scale of the seed train (80 L) using PLSR and SVR approaches.

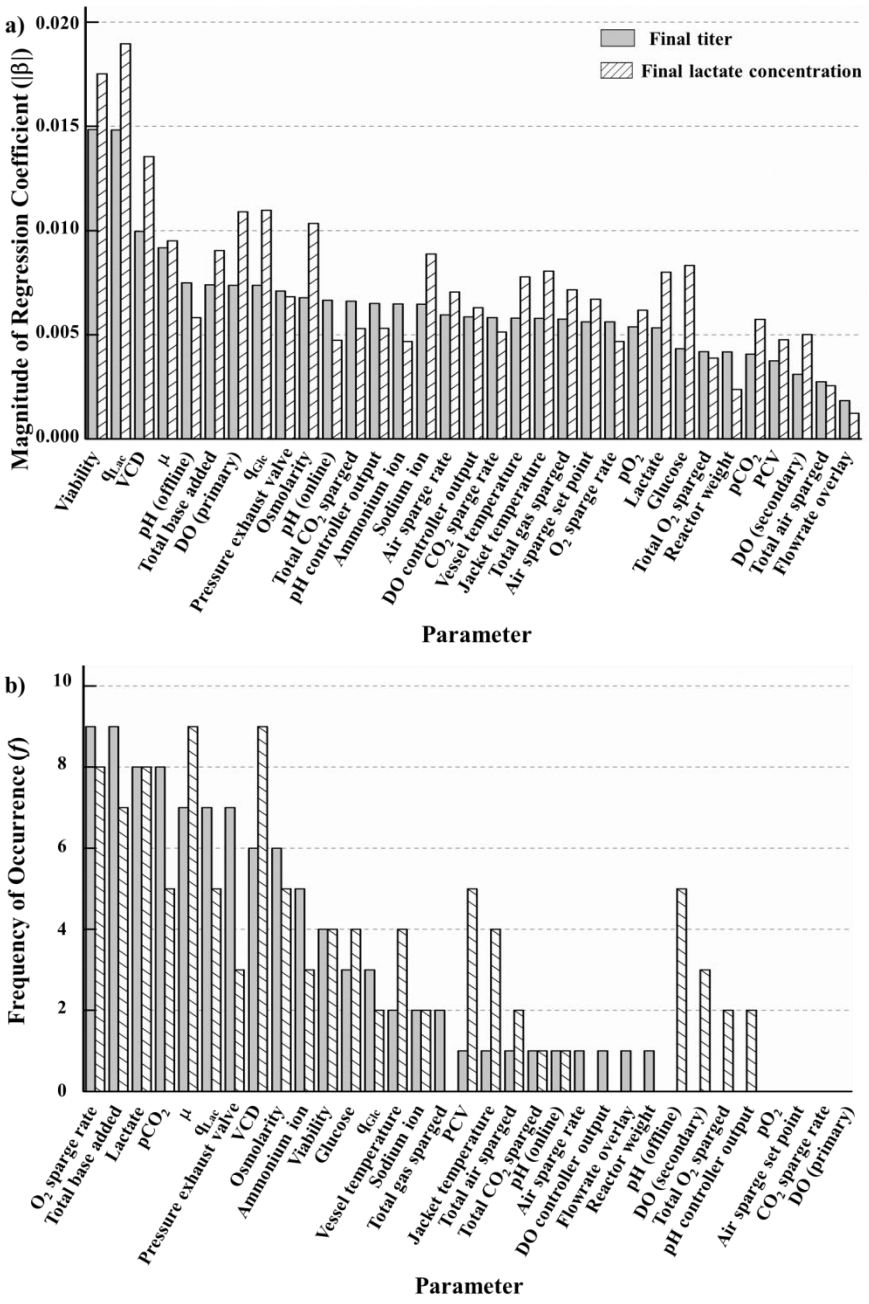


Figure 21: Contribution of process parameters to prediction accuracy of the final titer (▭) and the final lactate concentration (▨) using data acquired at 80 L scale bioreactors as evaluated using:

- **a) Magnitude of regression coefficient ($|\beta|$) of each parameter in optimized PLSR models**
- **b) Frequency of occurrence (f) of each parameter in optimized SVR models**

Both criteria of β and f led to a common conclusion that the majority of parameters pivotal to prediction of the final titer appeared to be related to cell growth and lactate metabolism by different degrees. These parameters include viable cell density, viability, specific cell growth rate, specific lactate production rate, total base added, lactate, and osmolarity. It is noteworthy that when the final lactate concentration was used as the objective function, similar parameters were identified as pivotal, supporting the notion that product yield and cellular metabolism are indeed strongly correlated.

The time profiles of several pivotal parameters in the top 20% and bottom 20% runs at the 80 L scale are shown in Figure 22. Although the differences between runs in these two classes were rather modest, a general trend could still be discerned. Runs in both classes appeared to be inoculated at similar cell concentrations, yet cells in most top 20% runs grew at relatively faster rates, giving rise to consistently higher viable cell density in these runs (Figure 22a and b). Cell viability was also largely remained high (> 90%) in these cultures (Figure 22c). Surprisingly, the majority of the top 20% runs experienced somewhat higher lactate concentration at this scale as shown in Figure 22d. However, the specific lactate production rate was lower (Figure 22e) and less base was added to maintain a constant pH (Figure 22f).

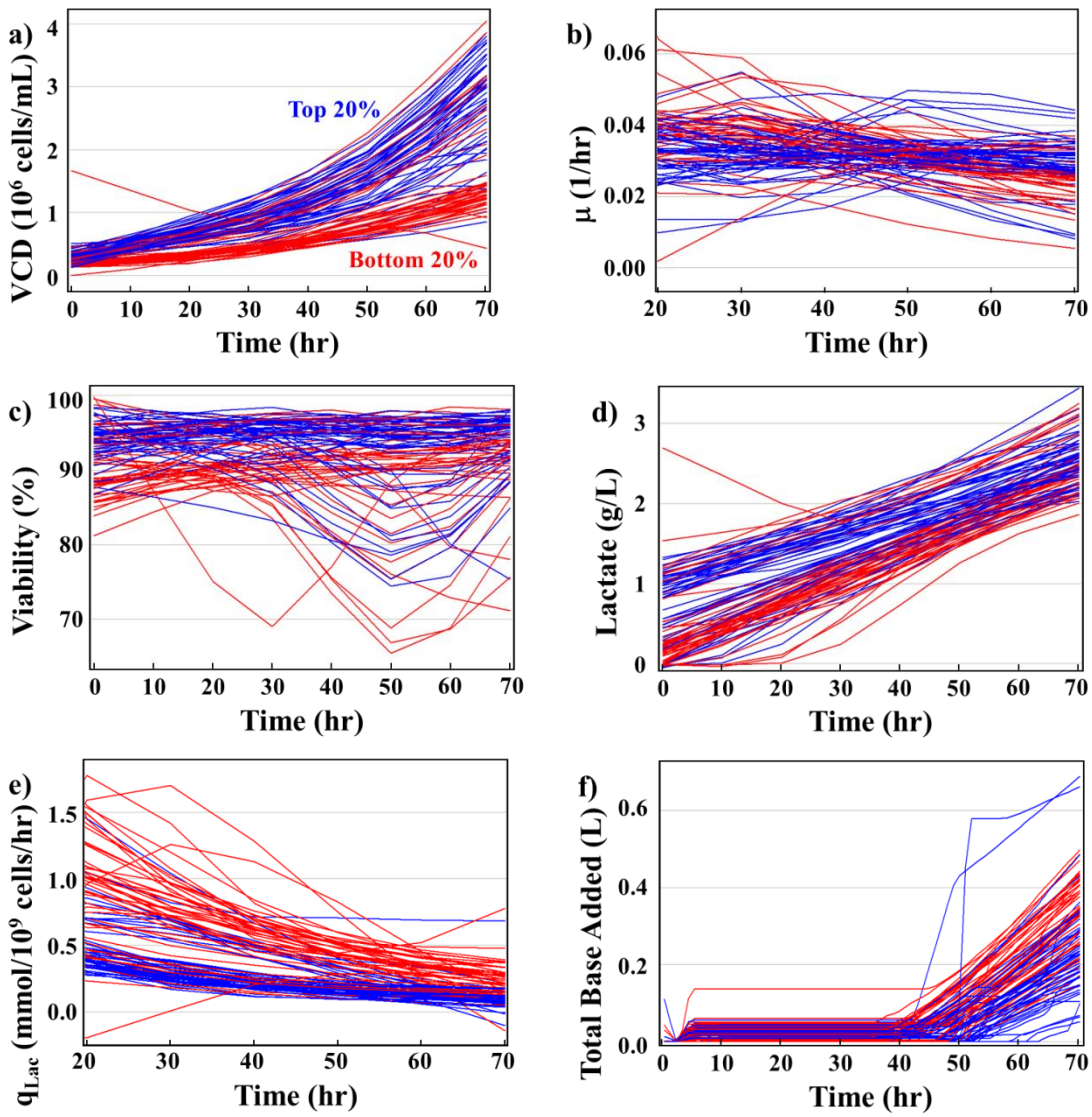


Figure 22: Time profiles of several pivotal parameters at 80 L scale.

Runs in the top 20% are colored in blue; those in the bottom 20% are in red.

- **a) Viable cell density**
- **b) Specific cell growth rate**
- **c) Viability**
- **d) Lactate concentration**
- **e) Specific lactate production rate**
- **f) Total base added**

It is interesting that these pivotal parameters identified during the seed train continued to be critical at the early stages of the production scale (Figure 23).

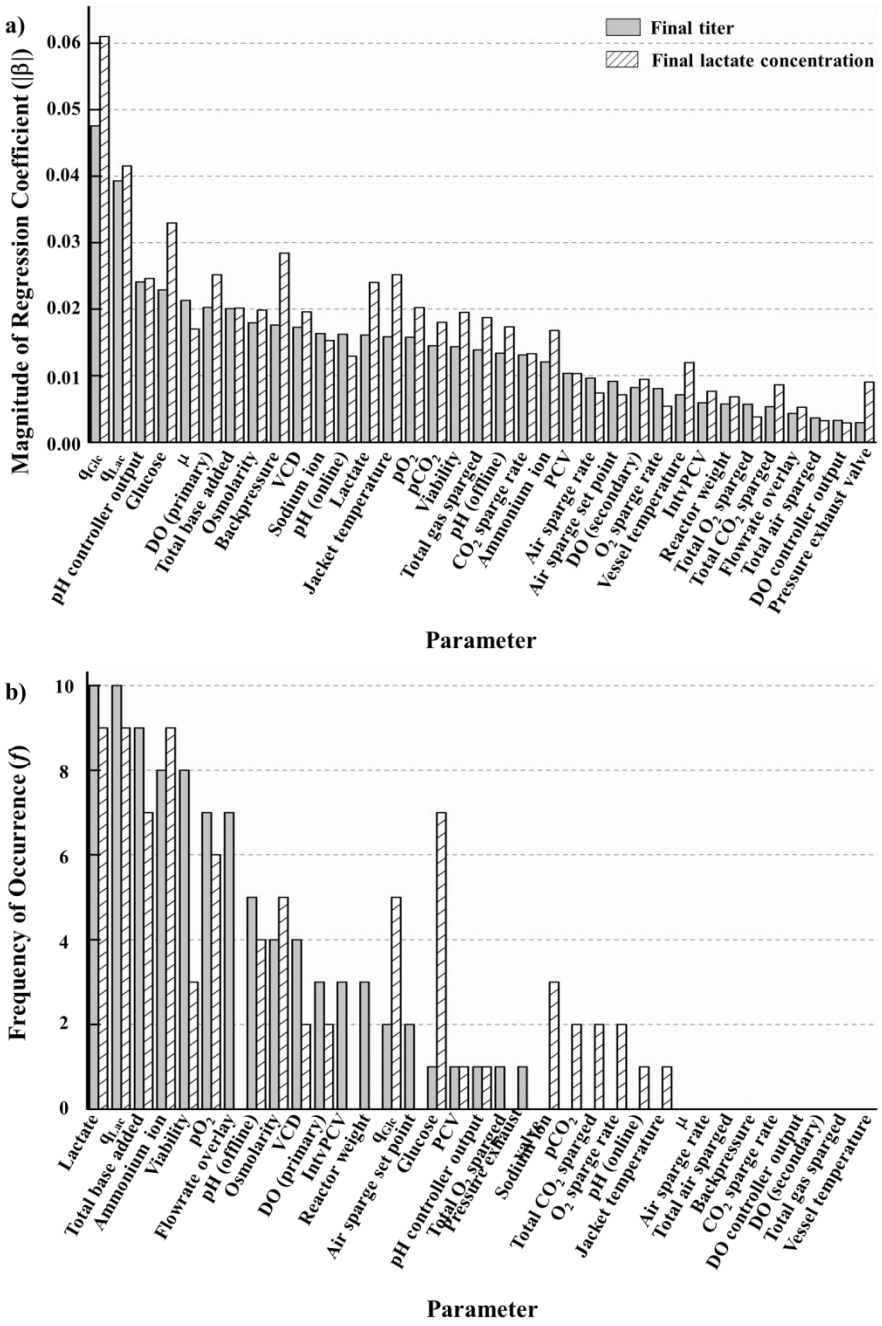


Figure 23: Contribution of process parameters to prediction accuracy of the final titer (▭) and the final lactate concentration (▨) using data acquired up to 70 hr of 12000 L scale bioreactors as evaluated using:

- **a) Magnitude of regression coefficient ($|\beta|$) of each parameter in optimized PLSR models**
- **b) Frequency of occurrence (f) of each parameter in optimized SVR models**

Furthermore, the subtle differences between runs in the top and bottom 20% classes observed at the 80 L scale were significantly magnified at the production scale (Figure 24).

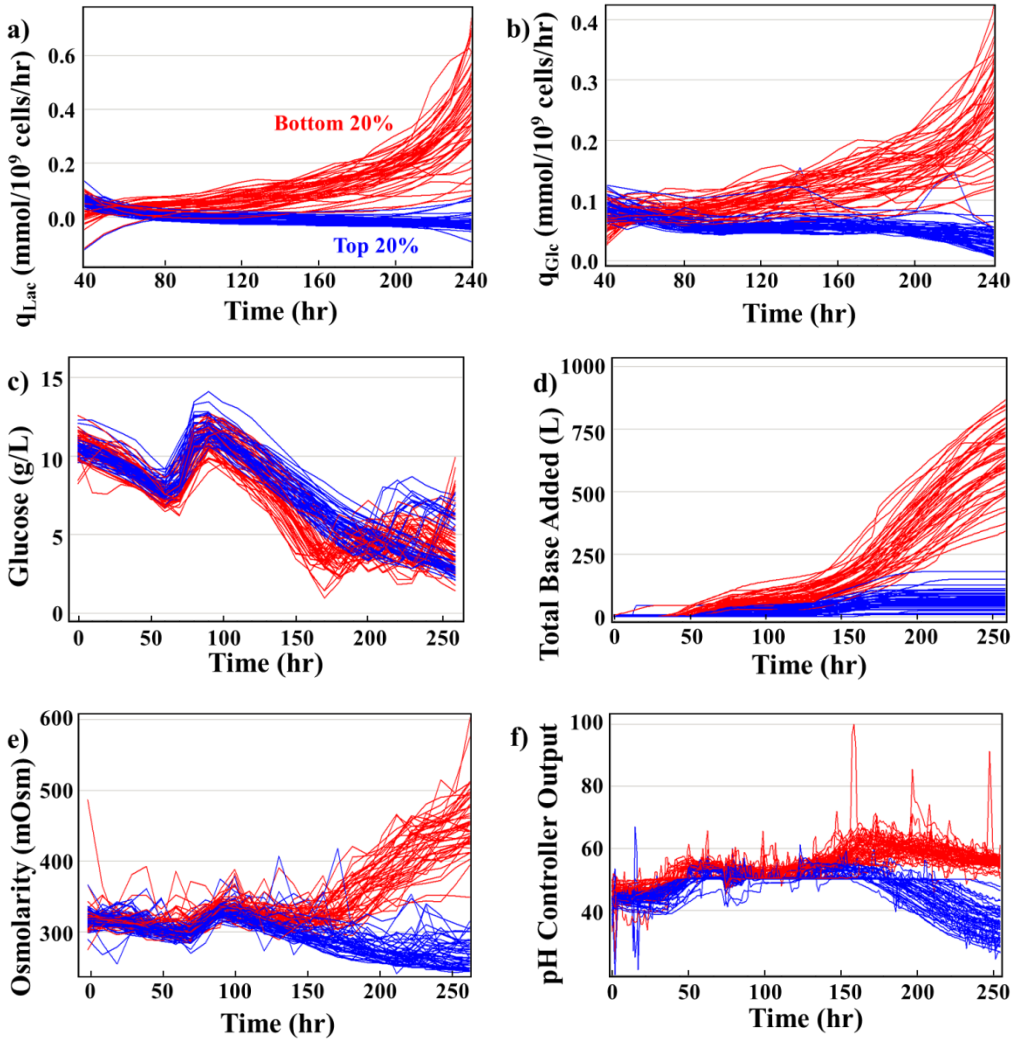


Figure 24: Time profiles of several pivotal parameters at 12000 L scale.

Runs in the top 20% are colored in blue; those in the bottom 20% are in red.

- **a) Specific lactate production rate**
- **b) Specific glucose consumption rate**
- **c) Glucose concentration**
- **d) Total base added**
- **e) Osmolarity**
- **f) pH controller output**

As early as 60 hr, a number of high-titer runs already had a metabolic shift to lactate consumption as indicated by negative specific lactate consumption rates, whereas most low-titer runs continued to produce lactate (Figure 24a). The majority of runs in the top 20% class eventually shifted to the lactate-consuming state. In contrast, runs in the bottom 20% class produced lactate at elevated rates, resulting in substantially high lactate concentrations in most cultures (Figure 18c).

Specific glucose consumption rates also differed significantly between the two classes (Figure 24b). High-titer runs consumed glucose at much reduced levels throughout the cultures than did those with low titer. Thus, high-titer runs did not appear to require multiple additions of glucose after the main feed at 70 hr (Figure 24c).

The low lactate concentration in the top 20% runs reduced or even eliminated the need for base addition whereas large amounts of base were added to the bottom 20% runs (Figure 24d). This base addition in turn led to accumulation of sodium ion to significantly higher concentrations (Figure 24e), and therefore osmolarity (data not shown), in these low-titer runs. The difference in lactate concentration between the two classes was also reflected in the pH controller output as shown in Figure 24f. The opposing behaviors of parameters related to lactate metabolism in the two classes further strengthened the findings that this set of parameters played an important role in predicting the final process outcome.

4.4.4 Lactate Consumption at Production Scale Emerges as Process Indicator

The analysis presented so far indicates a high correlation of cell growth and lactate metabolism to the final titer. The majority of parameters identified as pivotal for prediction of the final process outcome, using data from the seed train or the early stage of the production scale, are related to cell growth and lactate metabolism (Figure 21 and Figure 23). Runs with high viable cell concentration and low final lactate concentrations or consumed lactate at the production scale yielded high levels of recombinant antibody (Figure 18c and Figure 24a). Runs with low lactate production rates and high cell growth rates at the beginning of the seed train often had high final titer (Figure 22). Indeed, when specific lactate production rate was plotted against viable cell concentration or

specific glucose consumption rate at 80 L scale (Figure 25), two clusters of the top and the bottom 20% runs could be seen, albeit with a high degree of overlap. These metabolic indicators of the final process outcome thus hint at possible means to intervene with the process as early as the seed stage.

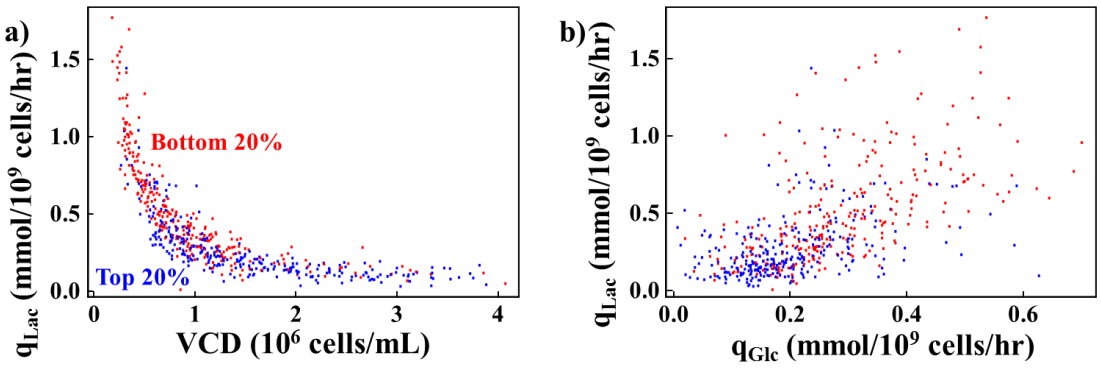


Figure 25: Relationship between several parameters related to cell growth and lactate metabolism for runs in the top 20% (blue) and the bottom 20% (red) classes at 80L scale.

Each data point represents one time point from 20 hr to 70 hr of 80 L cultures with 10 hr intervals.

a) Specific lactate production rate (q_{Lac}) vs. viable cell density (VCD)
b) q_{Lac} vs. specific glucose consumption rate (q_{Glc})

To gain more insights into the metabolic shift occurring at the production scale, which is highly correlated to hyper-productivity, the specific rates of lactate production, glucose consumption, and cell growth in the top and bottom 20% of runs at the late stage of the production scale (from 120 hr to 240 hr with 10 hr intervals) were further analyzed as shown in Figure 26.

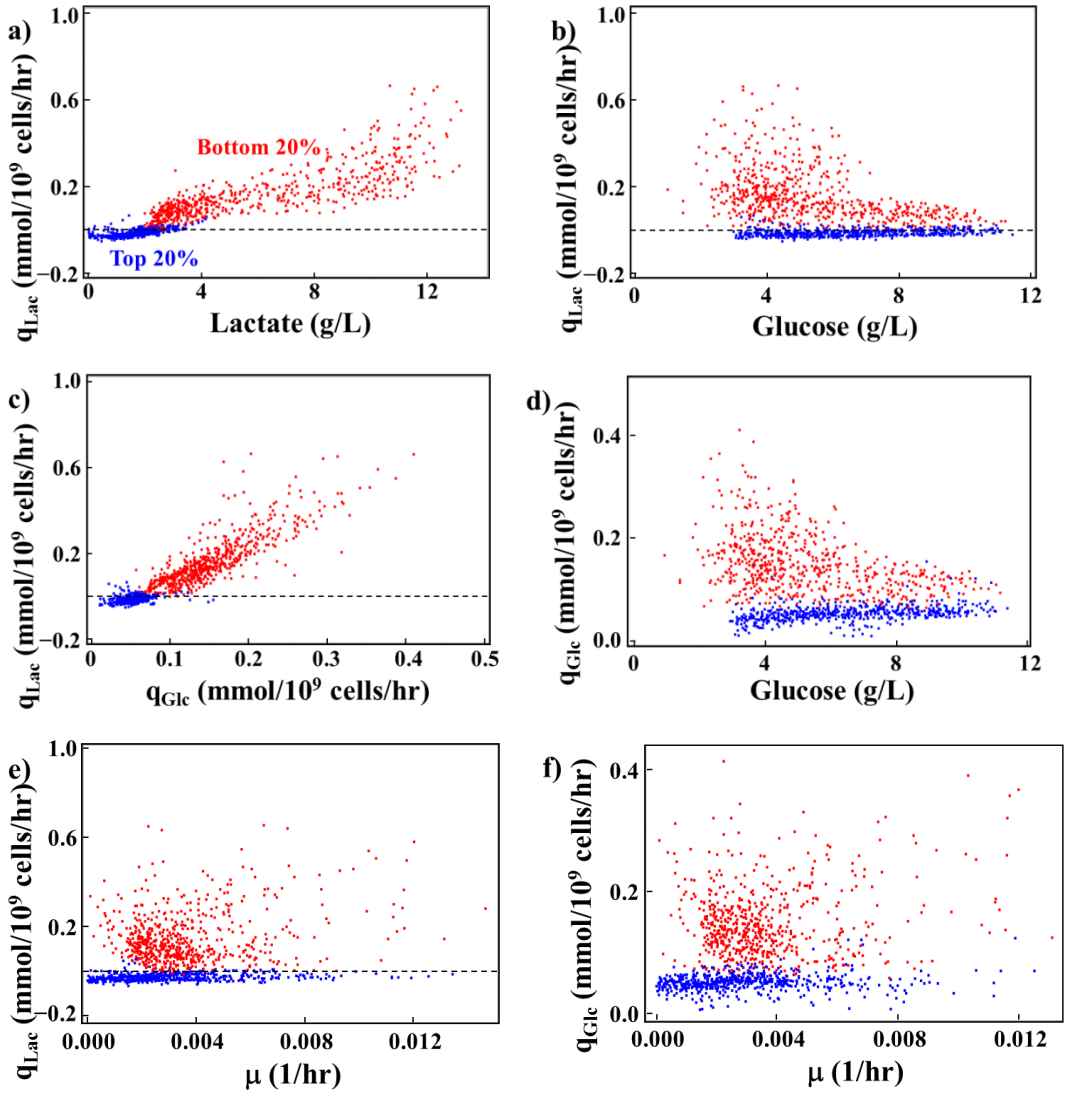


Figure 26: Relationship among several parameters related to cell growth and lactate metabolism for runs in the top 20% (blue) and the bottom 20% (red) classes in the late stage of the production scale.

Each data point represents one time point from 120 hr to 240 hr of 12000 L cultures with 10 hr intervals. The dashed line represents $q_{Lac} = 0$.

- **a) Specific lactate production rate (q_{Lac}) vs. lactate concentration**
- **b) q_{Lac} vs. glucose concentration**
- **c) q_{Lac} vs. specific glucose consumption rate (q_{Glc})**
- **d) q_{Glc} vs. glucose concentration**
- **e) q_{Lac} vs. specific cell growth rate (μ)**
- **f) q_{Glc} vs. specific cell growth rate (μ)**

In low-titer runs, specific lactate production rate spanned over a wide range from a very low value to as high as 0.6 mmol/10^9 cells/hr (Figure 26a). In contrast, specific

lactate production rate in high-titer runs spanned a much narrower range from 0.05 to – 0.05 mmol/10^9 cells/hr (consumption). Lactate consumption at the production scale, strikingly, occurred even when lactate was almost depleted in the cultures. This suggests that once cells start to consume lactate, they have a propensity to continue consuming it regardless of the low level of this metabolite. Likewise, cells in a lactate-producing culture appeared to remain in that state despite the extensive accumulation of lactate. In other words, high concentration of lactate alone is not sufficient to trigger lactate consumption, nor does it completely inhibit lactate production.

Interestingly, glucose concentration does not dictate lactate metabolism; both lactate production and consumption can occur over the same wide range of glucose concentration (Figure 26b). In other words, the abundant presence of glucose does not deter lactate consumption. It is evident that lactate consumption occurs only when the specific glucose consumption rate is low (below 0.07 – 0.1 mmol/10^9 cells/hr) (Figure 26c). There also seems to be a minimum specific glucose consumption rate that cells sustain, as the value never reaches zero. Furthermore, glucose concentration alone does not determine glucose consumption, as can be seen in Figure 26d. There is virtually no difference in the range of glucose concentration between metabolically shifted cultures ($q_{Lac} \leq 0$ and low q_{Glc}) and "typical" cultures (high q_{Lac} and high q_{Glc}). It should be noted that the glucose concentrations in all cultures were maintained at more than 3 g/L, substantially higher than the reported K_m of the GLUT1 transporter for glucose (approximately 0.18 g/L).

It is interesting to observe that lactate consumption did not depend on how much cell growth slowed down during the late stage of the production scale (Figure 26e). Both lactate-consuming and lactate-producing cultures appeared to have similar ranges of specific cell growth rate. Likewise, glucose consumption rate can vary vastly regardless of specific cell growth rate (Figure 26f). Taken together, these observations indicate that the potential of cells to consume lactate in the late stage is largely a function of reduced glycolytic flux rather than of glucose or lactate concentration or cell growth.

4.5 DISCUSSION

The immense volume of cell culture bioprocess data in historical archives certainly holds valuable insights into manufacturing processes and product characteristics. This resource has begun to be explored to generate process insights using multivariate data analysis tools in recent years. This study employed two such tools, SVR and PLSR, to investigate process data from more than two hundred production-scale cultures. Both methods could predict process performance with similar high accuracies if data from the production bioreactors were used with the objective function being either the final titer or the final lactate concentration. Data acquired at the seed train alone (80, 400 and 2000 L reactors) were somewhat less predictive of the final process outcome compared to the production-scale data. However, only a few of the predicted top 20% runs switched class to the bottom 20% in the actual outcome, even using only data from the seed train. Such lack of switching also holds true for the bottom 20% runs.

The pivotal parameters identified both at the seed train and at the production scale are mostly associated with cell growth and lactate metabolism, indicating the prominent role of cellular metabolism in determining product titer. Previous analysis of a subset of runs used in this study (Charaniya et al. 2010) and data from another manufacturing process (Kirdar et al. 2008) also led to a similar conclusion. The results from this study further indicate that lactate consumption at the production scale serves as an indicator of high productivity. However, the conditions that induce lactate consumption in the high titer runs at this scale are still unknown.

The pivotal parameters identified in the seed train possibly impart a longer lasting effect on the process outcome. Such potential occurrence of a "memory" effect reiterates our previous findings (Charaniya et al. 2010) and the results from another study (Ündey et al. 2010). This effect also has a strong implication for process operation during early stages as it appears to have a profound impact on the final process outcome. However, it is also important that the predicted class does change in some cases as evident in Figure 19e. For some runs, even though the pivotal process parameters (indicators) indicated that their process characteristics placed them in the low-titer class, they gradually switched to the high-titer class. Since those pivotal parameters are metabolism related, it

thus indicates that metabolic characteristics can be altered, but the alteration is relatively gradual. A better understanding of the relationship between the culture environment and metabolic characteristics may enlighten our approach of intervention. Or for the cases that the class switched from high-titer to low-titer, a better understanding of prevention of such a shift will be critical.

A key correlated factor of low lactate production and lactate consumption at the production scale appears to be low specific glucose consumption. Thus controlling glycolytic flux seems to be the key to modulating lactate metabolism and therefore the final product yield. Such a conclusion has also been reached through a metabolic study in conjunction with modeling (Mulukutla et al. 2011), which showed that the switch to the lactate consumption mode could be attributable to a moderate attenuation of glycolytic genes' expression and differential activities of the Akt and p53 signaling pathways. Indeed, inhibition of the Akt pathway by addition of its inhibitors in the late growth stage was shown to facilitate lactate consumption.

The analysis presented in this work may hint at possible routes of intervention to steer the low titer runs to higher productivity. One possible approach is by early intervention, as metabolic indicators at even the 80 L scale show a correlation to the final titer. An inverse relationship between viable cell concentration and specific lactate production, as well as a positive correlation between glucose consumption and lactate production, point to the possible benefit of rearranging seed train scheduling strategy and of selective use of 80 L runs for subsequent inoculation into 400 L runs.

Since the predictability at 80 L scale is only tentative, some production runs will inevitably have a propensity towards becoming low performers. Remedial corrective measures at the production scale will need to focus on manipulating cell metabolism prior to 70 hr, the point at which the correlation between predicted and actual titer still hints at some flexibility in the outcome. Many possible approaches of suppressing glucose metabolism and eliciting metabolic shift to lower lactate production or lactate consumption have been reported, including reducing glucose concentration (Cruz et al. 1999; Zhou et al. 1997), employing alternative sugars (Altamirano et al. 2006; Wlaschin and Hu 2007a), supplementing copper ion (Qian et al. 2011), and adding inhibitors of the

Akt pathway (Mulukutla et al. 2011). Conceivably interventive measures can be taken by then if necessary. Whether those possible interventions will be effective can only be answered by experimentation. Whether the intervention methods, if proven effective, should be implemented in a manufacturing setting will largely depend on the operating protocols of each individual plant and the nature of the interventions.

With the increasing emphasis on the concept of Quality by Design (QbD) in the production of therapeutic biologics, we foresee such practices of mining bio-manufacturing data being extended to analyze Critical Quality Attributes (CQAs). Recently, clustering of glycosylation profiles of an antibody product has revealed a high correlation between product quality attributes and process characteristics (Le et al.). As both process and product quality data continue to accumulate, the likelihood of identifying process characteristics which affect product quality will also increase. Harnessing the power of data mining will greatly strengthen our capability to produce high quality products through high-productivity processes.

5 Analysis of Transcription Dynamics of Selection and Gene Amplification

5.1 Summary

The dihydrofolate reductase (DHFR)-based amplification system is the most commonly used method in CHO cells for increasing the expression level of a product gene following selection for production of recombinant protein therapeutics. This approach has enabled the transformation of host cells from non-producers to professional secretors. However, three decades after the advent of recombinant DNA technology, we still have little understanding of the mechanisms of this transformation process.

To gain mechanistic insights into this process, a parental CHO cell line deficient in DHFR activity was transfected with a vector expressing a mouse dihydrofolate reductase (mDHFR) gene, a hygromycin resistance marker (HPT), and transgenes encoding for a human immunoglobulin G (hIgG) antibody product. A control experiment using a similar vector without the antibody transgenes was also carried out. Following transfection and initial selection in hygromycin, methotrexate (MTX) treatment was applied. Single clones at different stages were isolated, and representatives of varying productivities were subjected to transcriptome analysis. Amplified populations were further sub-cloned, and samples from isolated single sub-clones were also characterized in fed-batch cultures.

The integration and amplification of the vectors under the pressure of hygromycin and MTX was confirmed by an increase in mRNA levels of the hIgG, the exogenous DHFR, and the HPT genes. Surprisingly, more profound transcriptional changes were seen upon selection than amplification. The transcript levels of these transgenes showed a tremendous surge upon selection and only a moderate or even no increase upon amplification. Functional analysis revealed several signaling pathways, aminoacyl tRNA biosynthesis, cell cycle, DNA replication, and amino acid metabolism which were commonly enriched upon selection and amplification. The results suggest that the development of cellular machineries required for hIgG production occurred even prior to amplification.

The notable clonal variability during selection and amplification continued to occur among different sub-clones during fed-batch cultures. Two classes of sub-clones regarding product titer and cell growth, namely high-producers with slow growth and low-producers with fast growth, were discerned. Among cellular functions enriched in the high-producers were protein processing, cell cycle, cytoskeleton function, ECM receptor interaction, ABC transporter, and several signaling pathways.

We hypothesize that the selection process, with the forced expression of a secretory protein hIgG, enriched survivors with superior secretion machineries. The role of the amplification process is less about further enhancement of the transgene expression level than further reinforcement of other cellular characteristics which favor high productivity, including protein processing, cell cycle, and signaling pathways. The mechanistic insights gained through such systems analysis will allow for a rational screening strategy of robust production cell lines.

5.2 INTRODUCTION

The past few decades have witnessed a tremendous increase of more than 100 fold in product titer achieved in recombinant protein-producing cells (Wurm 2004). In addition to downstream optimization, this achievement can be attributed to relentless improvement in the most upstream process of cell line development, which includes product gene delivery, selection, amplification, and screening. The transformation of protein secretion capability of these host cells during the cell line development process, from none to levels which rival professional secretors *in vivo*, entails profound changes in gene regulation which are poorly understood. Empirical selection of cells with high productivity and desired growth characteristics has been the key to this success. However, as the demand for recombinant protein therapeutics continues to grow, it is necessary to gain better understanding into the physiological mechanisms underlying this transformation process.

Typically, a plasmid DNA molecule(s) carrying a product gene and a selection marker can be introduced into the host cells using a multitude of DNA delivery methods such as calcium phosphate, electroporation, lipofection, and polymer-mediated gene

transfer. Subsequently, this exogenous DNA fragment would integrate into the host genome, by and large randomly via chromosomal rearrangement and non-homologous recombination, providing cells with a survival advantage under the selection pressure (Finn et al. 1989). Due to variability in integration site and plasmid copy number, substantial screening is often performed following selection to isolate single clones with exceptional growth characteristics and titer.

To further increase productivity, the selected clones are often subjected to gene amplification, which co-amplifies the product gene along with the selection marker. The most commonly used amplification system is dihydrofolate reductase (DHFR) and its chemical antagonist, methotrexate (MTX) (Bendig 1988; Gandor et al. 1995; Pallavicini et al. 1990). By using a host cell line deficient in DHFR activity and by introducing an exogenous DHFR alongside the product gene, one can achieve co-amplification via adding MTX at increasing concentrations. It has been shown that MTX-mediated gene amplification could result in hundreds to thousands copies of the plasmid in tandem repeats in the host genome (Wurm et al. 1986). Thereby, specific productivity could increase up to 10- to 20-fold (Wirth et al. 1988).

Both selection and amplification processes, upon which the host cells are completely transformed from non-secretors to professional secretors, are likely to confer drastic changes in multiple cellular functions. These changes are essential to cope with the expansion of the transcription, mRNA processing, translation, protein processing, and secretory machineries. However, there has been a lack of studies investigating the transcriptional changes in the host cells undergoing such transformations. Such understanding could be achieved at the transcriptome level through the use of DNA microarrays, providing a mechanistic guide to the screening process.

In this study, we isolated a panel of hIgG-producing clones through transfection with both chains of the hIgG gene, the hygromycin resistance marker (HPT), and the DHFR gene. Following selection in hygromycin, these clones were further subjected to MTX-mediated amplification. Transcriptome analysis using microarrays was performed on each stage of the process: host, post-selection, and post-amplification. This study therefore represents the first example of applying transcriptome analysis to study the

mechanisms underlying selection and amplification. Furthermore, several single sub-clones spanning a wide range of titer were generated from the amplified hIgG-producing populations. Samples from day 4 and day 7 of fed-batch cultures of these sub-clones were also analyzed using microarrays. As a result, a set of genes potentially conferring hyper-productivity traits could be compiled from these analyses and several previous studies.

5.3 MATERIALS AND METHODS

5.3.1 Generation of hIgG-producing Clones and Non-producing Pool

CHO DXB-11 cells (host, H) were cultured in MEMα+HT medium supplemented with 10% fetal bovine serum (FBS) at 37^{o}C and 7.5% CO_2 in a humidified environment. Cells at 80% confluency in T75 flasks were transfected with 4 μg each of two expression vectors, pHC and pLC, using Lipofectamine 2000 (Invitrogen, Carlsbad, CA). The pHC vector carries a hIgG heavy chain driven by a CMV promoter and a hygromycin resistance marker (HPT) driven by a TK promoter. The pLC vector contains a hIgG light chain driven by a CMV promoter and an mDHFR gene driven by an LTR promoter.

Forty-eight hours following transfection, the cells were diluted in 96-well plates in MEMα supplemented with 10% FBS and 400 μg/mL hygromycin. After three weeks, ELISA was used to measure hIgG concentration (titer) in the supernatant of isolated single clones. The three best clones in terms of titer, namely P_1, P_2, P_3, were selected for further analysis.

In addition to these hIgG-producing clones, a transfection with a control vector carrying only the HPT and the mDHFR genes was performed in parallel. The transfected cells were selected and expanded in T75 flasks as a control pool (C). The absence of hIgG heavy chain and light chain in this control pool allowed us to evaluate the effect of the hIgG genes alone.

5.3.2 MTX-based Gene Amplification

The three hIgG-producing clones and the control pool generated as described above were treated with 20 nM of MTX (Sigma Aldrich, St. Louis, MO) for 15 days in

duplicate cultures in T75 flasks. Cell samples were taken immediately before MTX treatment ($P_{1,0}$, $P_{2,0}$, $P_{3,0}$, and C_0) and after MTX treatment ($P_{1,M}$, $P_{2,M}$, $P_{3,M}$, and C_M) for microarray analysis. In addition, duplicate samples of the host cells (H) were also analyzed. The experiment design is briefly illustrated in Figure 27.

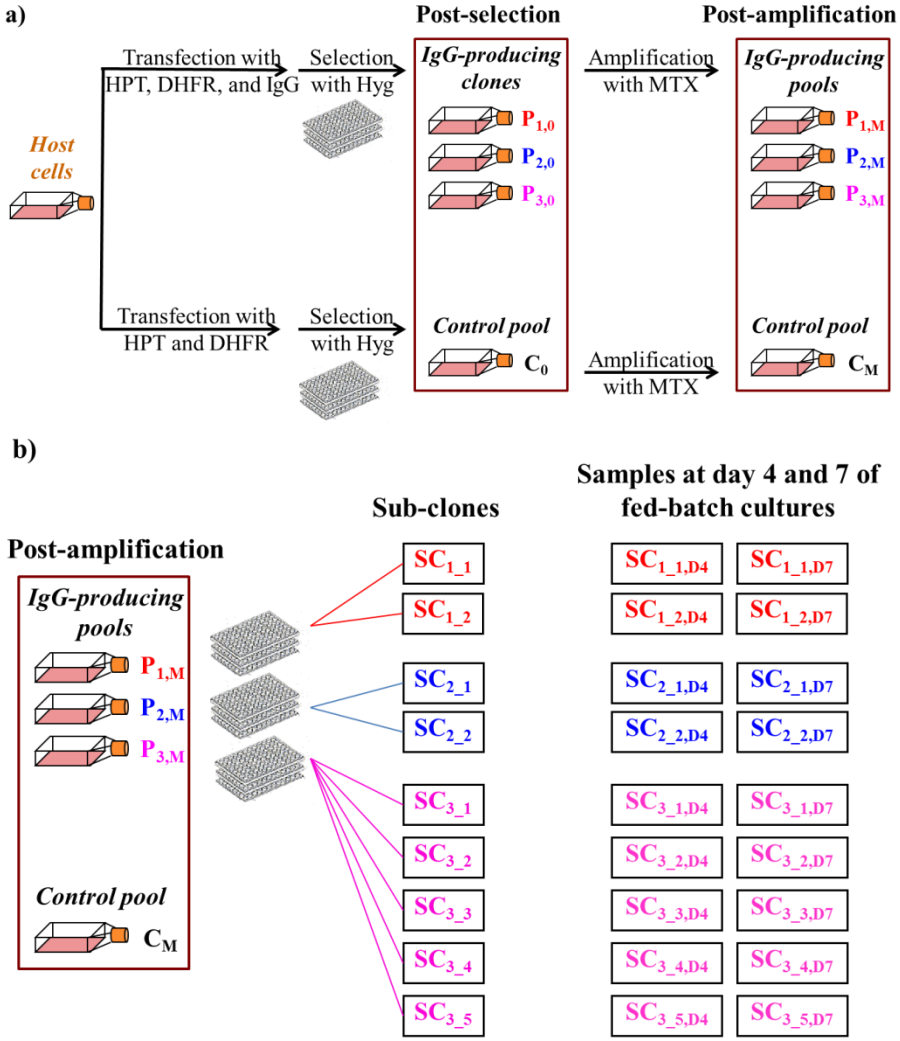


Figure 27: Experiment design of selection, MTX-mediated amplification, and fed-batch cultures.

a) Host cells were transfected with hIgG, mDHFR, and HPT. The transfected cells were selected in hygromycin and single cell cloned. The three best clones were picked and subjected to MTX treatment for 15 days. A control pool was generated through a similar process.
b) Amplified hIgG-producing populations were further sub-cloned. Nine sub-clones were picked and used to run fed-batch cultures. Duplicate samples of host, hIgG-producing clones and control pool before and after amplification, and nine sub-clones at day 4 and day 7 of fed-batch cultures, were analyzed using Affymetrix microarrays.

Following MTX treatment, the three hIgG-producing populations were sub-cloned in 96-well plates. A panel of nine sub-clones (SC) spanning a wide range of hIgG concentrations in the supernatants was selected. These sub-clones were adapted to growth in suspension prior to inoculation of fed-batch cultures at 2 L scale. Cell lysates collected on day 4 and day 7 of these sub-clones were used for microarray analysis.

5.3.3 Transcriptome Analysis

Total RNA was isolated using the RNeasy Mini kit (Qiagen, Valencia, CA) according to the manufacturer's protocol with on-column DNase I treatment for removal of genomic DNA. First-strand and second-strand synthesis and labeling were performed using 5 μg of total RNA following a protocol recommended by Affymetrix (Santa Clara, CA). Ten micrograms of labeled, fragmented cRNA was hybridized onto custom Affymetrix microarrays with 61,223 probe sets. Microarrays were washed using the Affymetrix GeneChip Fluidics Station 450 and scanned using the Affymetrix GeneChip Scanner 3000.

For mDHFR and HPT, which were not represented by any probe sets on the array, quantitative real-time PCR (qRT-PCR) was used to measure mRNA levels. First-strand cDNA was synthesized from 2.5 μg of total RNA using Superscript III Reverse Transcriptase (Invitrogen, Carlsbad, CA) with 1 mM oligo dT primers in a total volume of 50 μL. A no reverse transcriptase control was performed in parallel to assess the extent of genomic DNA contamination. The primers used for qRT-PCR are listed in Table 3.

Table 3: List of qRT-PCR primers used for quantification of mRNA levels of mDHFR, HPT, and β-actin.

Gene	Left primer	Right primer
mDHFR	TCTGTTTACCAGGAAGCCATGA	AATTCCTGCATGATCCTTGTCA
HPT	GATGTTGGCGACCTCGTATT	GATGTAGGAGGGCGTGGATA
β-actin	GTCGTACCACTGGCATTGTG	AGGGCAACATAGCACAGCTT

The mRNA levels of these two genes were quantified using the Brilliant II SYBR Green qPCR Master Mix (Agilent, Santa Clara, CA). Each 12.5 μL reaction mixture contained 6.25 μL of the master mix, 50 ng of the cDNA template, and 0.2 μM of each primer (forward and reverse). qRT-PCR was performed using the Stratagene Mx3000P instrument (Agilent, Santa Clara, CA). The thermal cycling profile was set as follows: initial denaturation at 95^{o}C for 10 min, followed by 40 cycles of 95^{o}C for 10 sec, 57^{o}C for 1 min, and 72^{o}C for 30 sec. The dissociation curves of the PCR products were generated by ramping from 57^{o}C to 95^{o}C after a denaturation step at 95^{o}C for 1 min and an annealing step at 57^{o}C for 30 sec. All cDNA samples were run in triplicates alongside a no reverse transcriptase control and a no template control. β-actin was used as a reference for comparison across samples.

5.3.4 Microarray Data Processing and Analysis

Expressionist software (GeneData, San Francisco, CA) was used to process all array image (.CEL) files. The raw intensities were condensed using the Microarray Analysis Suite (MAS 5.0) Statistical Algorithm (Affymetrix, Santa Clara, CA). After background removal, a linear normalization was performed to scale the average intensity value of all probe sets within each array to 500. Probe sets with a minimum detection *p*-value ≥ 0.04 or a maximum intensity value ≤ 70 in all samples were called absent and removed from further analysis.

Hierarchical clustering of microarray data was performed using Spotfire DecisionSite 9.1.2 (Somerville, MA). Default settings of the distance metrics (Euclidean) and the clustering method (unweighted average) were used.

Differentially expressed genes were identified using Significance Analysis of Microarray (SAM) (Tusher et al. 2001). Genes with a fold-change ≥ 1.5 and a q-value $\leq$ 10% were considered statistically differentially expressed.

Functional analysis was performed using Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005). Approximately 550 curated gene sets from the Broad Institute were used (http://www.broadinstitute.org/gsea/downloads.jsp). Gene sets with a nominal *p*-value ≤ 0.07 were considered significantly enriched.

5.4 RESULTS

5.4.1 Profound Transcriptional Changes of Transgenes Imposed by Selection and Amplification

A panel of three hIgG-producing clones (P_1, P_2, P_3) and a control pool (C) were subjected to MTX treatment following selection in HT-minus medium supplemented with hygromycin (Figure 28a). Duplicate samples were taken for transcriptome analysis immediately before MTX treatment (post-selection) and 15 days following treatment (post-amplification). Host cells were also sampled in duplicate to serve as a baseline for evaluating the magnitude of transcriptional responses induced by selection and amplification.

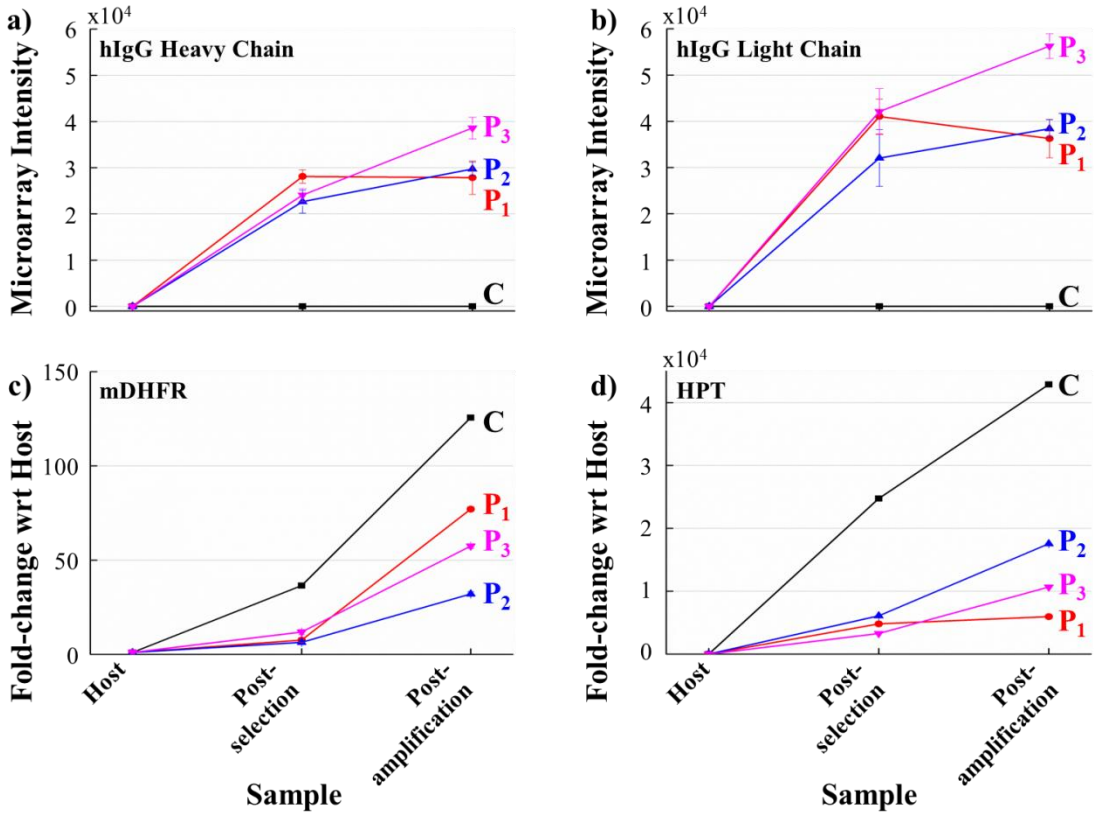


Figure 28: Changes in mRNA levels of transgenes in hIgG-producing clones and control pool after selection and amplification compared to host cells.

- **a) hIgG heavy chain (microarray intensity)**
- **b) hIgG light chain (microarray intensity)**
- **c) mDHFR (fold-change with respect to host using qRT-PCR)**
- **d) HPT (fold-change with respect to host using qRT-PCR)**

As evident in Figure 28a, all three hIgG-producing clones displayed a tremendous surge in hIgG heavy chain mRNA level. After selection, this level increased from almost none in host cells to approximately 20,000-30,000 of normalized microarray intensity. Note that the average intensity of all genes in an array is 500. The change in hIgG heavy chain expression level following amplification, surprisingly, was rather modest. In clone 1 (P_1), this level stayed relatively constant at 28,000. In clones 2 (P_2) and 3 (P_3), a 1.3-1.6 fold increase to 30,000-40,000 was observed. As a side note, no signal of hIgG heavy chain was detected in the control pool (C), indicating that the probe set designed for hIgG heavy chain on the Affymetrix microarray was specific.

hIgG light chain expression level also experienced a similar trend following selection and amplification (Figure 28b). After a surge from almost zero to 30,000-40,000 of normalized microarray intensity upon selection, the expression level of hIgG light chain in P_2 and P_3 moderately increased 1.2-1.3 fold upon amplification. P_1 even experienced a slight decrease in hIgG light chain expression level of 1.1 fold.

As mDHFR and HPT are not represented by any probes on the array, qRT-PCR was performed and a fold-change with respect to the host cells was reported. The change in mDHFR mRNA level upon selection and amplification appeared to be more consistent, albeit slightly smaller in magnitude (Figure 28c). The DHFR expression level was approximately 6-12 fold higher in the three clones post-selection compared to that in the host cells. Upon amplification, the gene underwent another 5-10 fold increase in expression. Thus, compared to the host cells, the amplified populations expressed DHFR 30-80 fold higher. The change of DHFR expression level in the control pool was more significant: 35-fold increase upon selection and a further 3.5 fold increase (to an overall fold-change of 125) upon amplification.

Compared to DHFR, the hygromycin resistance marker (HPT) displayed a more drastic change following selection and amplification (Figure 28d). All selected clones expressed HPT at a level 3,000-6,000 fold higher than the host cells. Up to 3-fold increase was further induced upon amplification, which led to a total fold-change of 6,000-17,000 in the amplified populations. In the control pool, the fold-change was 25,000 upon selection, and 43,000 upon amplification.

In most cases, selection surprisingly appeared to induce more drastic change in the expression levels of the four transgenes than did amplification. The fold-change upon selection could be tens of thousands, whereas the change upon amplification was limited below ten.

5.4.2 Global Transcriptional Changes Following Selection and Amplification

To characterize transcriptional dynamics in response to selection, we compared the expression levels in post-selection samples of hIgG-producing clones (P_1, P_2, and P_3) and control pool (C) to host cells. As shown in Table 4, in P_1, P_2, and C, a large number of genes (on the order of 700 – 900) were considered differentially expressed upon selection with a fold-change ≥ 1.5 and a q-value $\leq 10\%$ using Significance Analysis of Microarray (SAM). This number was rather small in P_3 – only 64 genes were differentially expressed.

When the same criteria were applied to compare samples before and after amplification, only about 200 – 500 genes were given differential calls in P_1, P_3, and C. Surprisingly, P_2 experienced a relatively modest change – only 22 genes changed their expression levels upon amplification.

Table 4: Number of differentially expressed genes in hIgG-producing clones and control pool upon selection and amplification.

Processes / Samples	Selection		Amplification	
	Up-regulated	**Down-regulated**	**Up-regulated**	**Down-regulated**
P_1	426	500	187	123
P_2	263	366	12	10
P_3	35	29	261	270
C	397	417	112	133
$P_1 \cap P_2 \cap P_3$	9	6	0	0

Functional analysis using Gene Set Enrichment Analysis (GSEA) was performed to identify cellular functions which are statistically enriched among all genes present on the microarray (Table 5). Unlike several other functional analysis approaches wherein

only differentially expressed genes were taken into consideration, GSEA is not affected by differential analysis results. Instead, it takes all genes into account by calculating a signal-to-noise ratio (Subramanian et al. 2005). Thus it allows small yet concerted changes within a functional pathway to be detected.

Gene sets enriched with a nominal p-value ≤ 0.07 in the selected clones compared to the host cells include several signaling pathways, cell cycle, DNA replication, mRNA processing, amino acid metabolism, amino-acyl tRNA biosynthesis, and glutathione metabolism. Of those, glutathione metabolism, mRNA processing, and a few signaling pathways were unaltered in the control pool.

Upon amplification, the EDG1, PDGF, and Ras signaling pathways, glycolysis, the TCA cycle, protein processing, amino acid metabolism, and amino-acyl tRNA biosynthesis were enriched in the amplified populations. Among them, glycolysis, the TCA cycle, and several signaling pathways were unchanged in the control pool.

Table 5: List of functional classes enriched in hIgG-producing clones and control pool upon selection and amplification.

Gene sets enriched upon selection	Gene sets enriched upon amplification
EDG1 signaling pathway*	EDG1 signaling pathway*
PDGF signaling pathway*	PDGF signaling pathway*
Aminoacyl tRNA biosynthesis	Aminoacyl tRNA biosynthesis
Cell cycle	Cell cycle
DNA replication	DNA replication
Amino acid metabolism	Amino acid metabolism
mTOR signaling pathway*	Ras pathway
Toll-like receptor signaling pathway	Extracellular matrix receptor interaction*
Cytokine-cytokine receptor interaction	Ribosome
mRNA processing*	Proteasome
Glutathione metabolism*	Glycolysis and TCA cycle*

* Unchanged in control pool.

5.4.3 Significant Clonal Variability upon Selection and Amplification

As seen in Figure 28 and Figure 29, all four transgenes were expressed at significantly different levels among the three hIgG-producing clones. P_3 appeared to consistently express both hIgG heavy chain and light chain at higher levels compared to P_1 and P_2 following amplification. For mDHFR and HPT, notable differences among the three clones were also observed although no consistent trend could be discerned. It is important to note that the difference in expression of the transgenes among these clones upon selection was rather modest. However, this difference appeared to be widened following amplification.

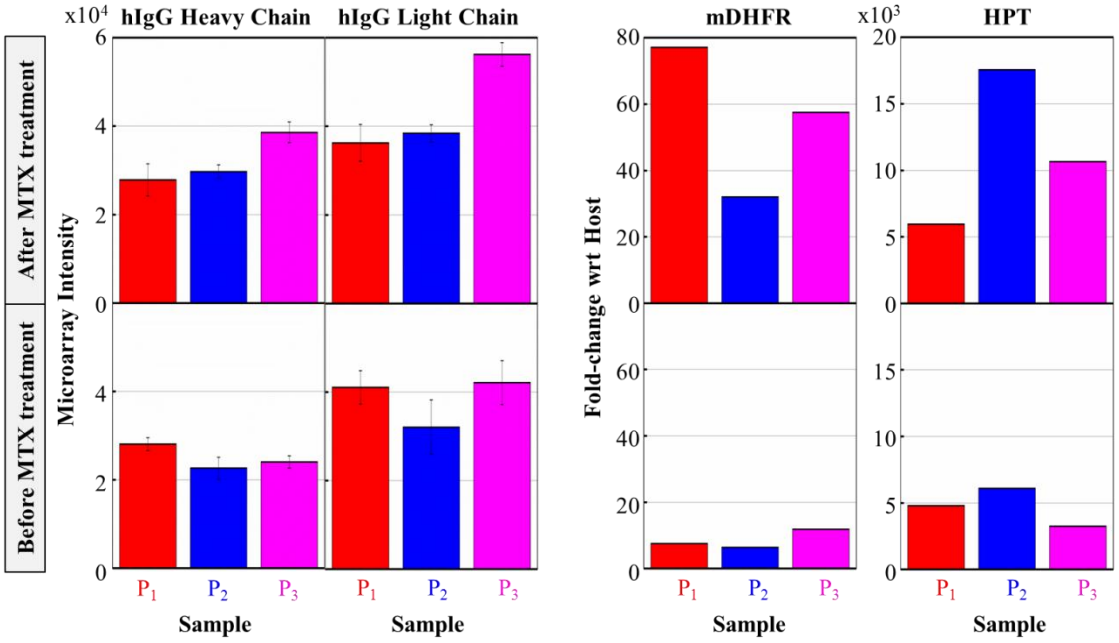


Figure 29: Variability among different clones in expression levels of hIgG heavy chain, hIgG light chain, mDHFR, and HPT before amplification (bottom panel) and after amplification (top panel).

In addition, as shown in Table 4, even though a large number of genes were differentially expressed following selection and amplification in each individual clone, the overlap among them seemed to be marginal. Instead of hundreds, only 15 genes were commonly differentially expressed among the three clones upon selection, and none upon amplification. This result indicated that these clones may have responded to the pressure of selection and amplification in considerably different ways.

Furthermore, hierarchical clustering was performed to assess the transcriptional similarity present amongst all analyzed samples (Figure 30). A greater degree of similarity was found within individual clones and the control pool, rather than before and after MTX treatment. In other words, amplification did not appear to greatly alter the transcription signature that had been established upon selection. This result suggests that each clone isolated following selection continued to carry distinct transcriptome characteristics over the course of MTX treatment.

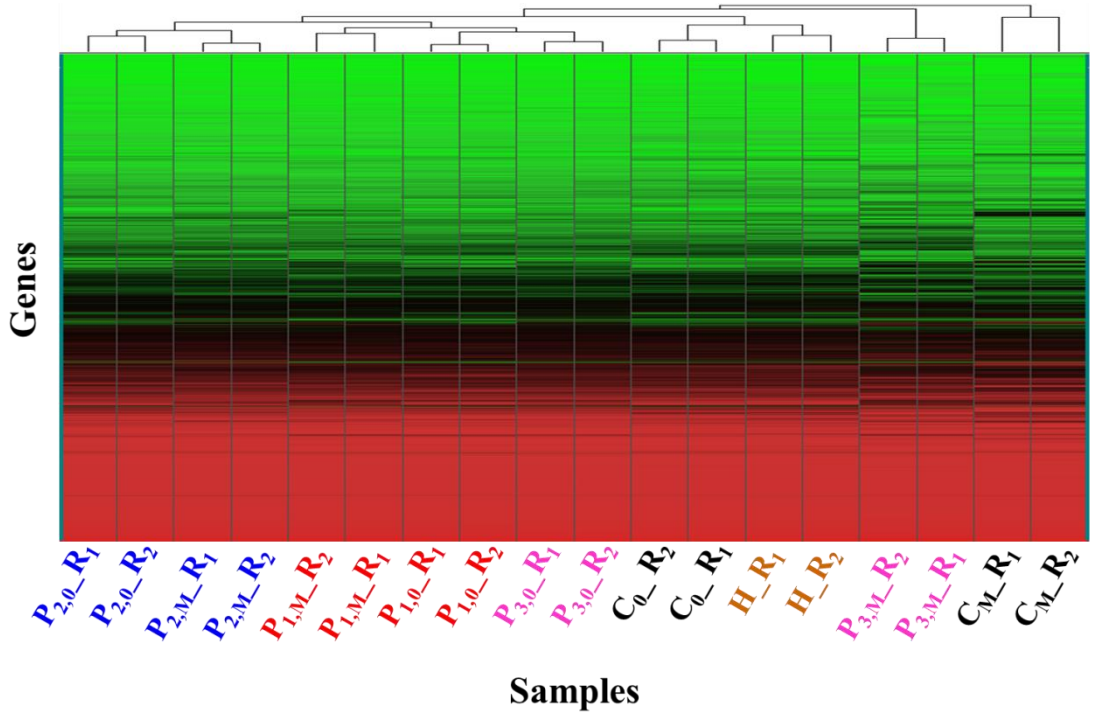


Figure 30: Variability among different clones revealed by hierarchical clustering of transcriptome data.

H: host, C: control pool, P: hIgG-producing clones, R: replicate, subscript zero: before MTX treatment, subscript M: after MTX treatment.

5.4.4 Wide Range of Titer and Growth Characteristics in Fed-batch Cultures

Each of the three hIgG-producing clones described above was further sub-cloned following amplification. Two sub-clones were picked from each of the first two clones (P_1 and P_2), and five sub-clones from the third one (P_3). Duplicated fed-batch cultures at 2 L scale were performed on the nine sub-clones. As shown in Figure 31, these sub-

clones produced hIgG at substantially different levels and displayed various growth characteristics. The final titer in these cultures varied within a wide range, from 200 to 900 mg/L (Figure 31a). The peak viable cell concentration (VCC) ranged from 6 to $11x10^6$ cells/mL (Figure 31b). It is interesting to note that the sub-clone with highest final titer (in blue) had the lowest peak VCC, and vice versa, the culture with highest peak VCC (in red) produced the lowest final level of hIgG. A similar inverse correlation could also be seen between the final titer and the final VCC. In addition to having higher peak and final VCCs, this low-producing sub-clone also maintained substantially higher viability during the first half of the fed-batch cultures (Figure 31c). However, shortly after the cells reached peak VCC at day 8, the viability in this culture dropped sharply. In contrast, the best producing sub-clone maintained a steady and slow rate of decreasing viability over the course of the fed-batch cultures, and thus ended with the highest final viability. The plot shown in Figure 31d indicated that the high titer cultures also had higher specific productivity. Thus all correlations derived above for high and low titer cultures could also be applied for cultures with high and low specific productivity.

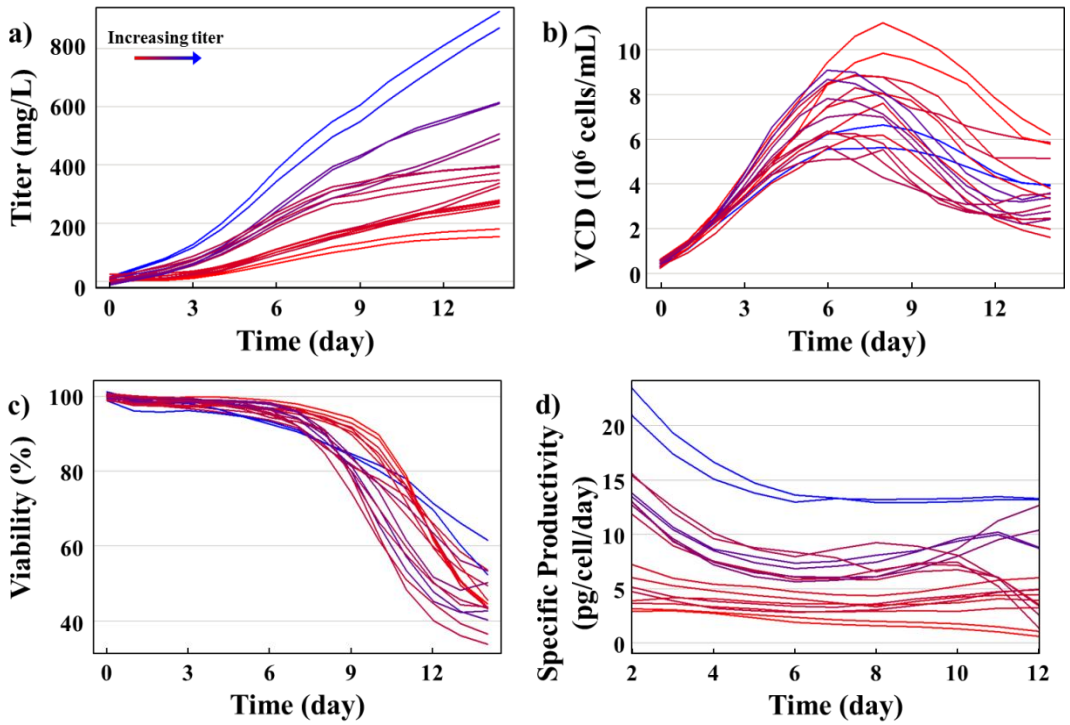


Figure 31: Various titer values and growth characteristics of sub-clones in fed-batch cultures.

The red to blue color gradient corresponds to low to high final titer.

Furthermore, the difference in product titer among sub-clones at day 4 and day 7 of the fed-batch cultures appeared to be well correlated to the difference in hIgG heavy chain mRNA levels (Figure 32). The correlation coefficient between them was 0.76 at day 4 and 0.69 at day 7. In contrast, the light chain mRNA level did not seem to dictate the titer, as their correlation coefficient was almost zero at day 4 and only 0.32 at day 7. Surprisingly, the DHFR mRNA level was inversely correlated to the hIgG titer. This correlation coefficient was – 0.47 and – 0.65 at day 4 and day 7, respectively.

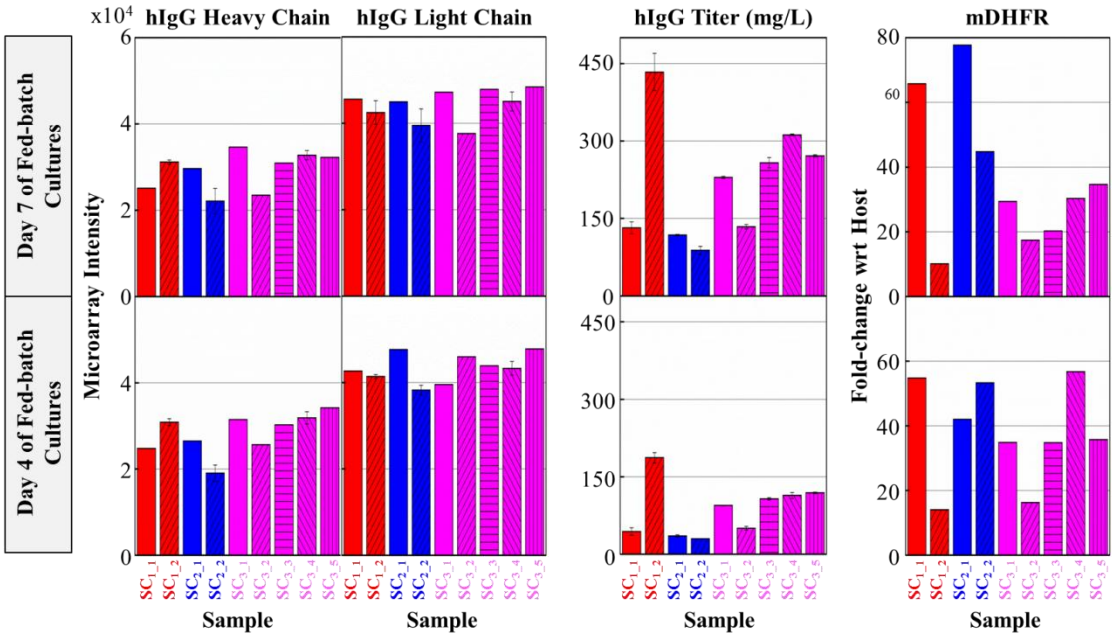


Figure 32: Variability amongst different sub-clones in expression levels of the transgenes and the hIgG titer at day 4 (bottom panel) and day 7 (top panel) of fed-batch cultures.

5.4.5 Transcriptome Analysis of High vs. Low IgG-producing Sub-clones in Fed-batch Cultures

Transcriptome analysis was performed on samples at day 4 and day 7 of fed-batch cultures to investigate the underlying transcriptional differences among the nine sub-clones. When the hIgG titers of these sub-clones at the two time points were re-arranged as shown in Figure 33, two classes of producers emerged: a high-producing class of five sub-clones (SC_{1_2}, SC_{3_4}, SC_{3_3}, SC_{3_5}, and SC_{3_1}) and a low-producing class of four sub-clones (SC_{1_1}, SC_{3_2}, SC_{2_1}, and SC_{2_2}). Subsequently, an approach similar to the one applied to analyzing transcriptional responses upon selection and amplification, including statistical analysis (SAM) and functional analysis (GSEA), was taken to explore the differences between the two classes at transcription level.

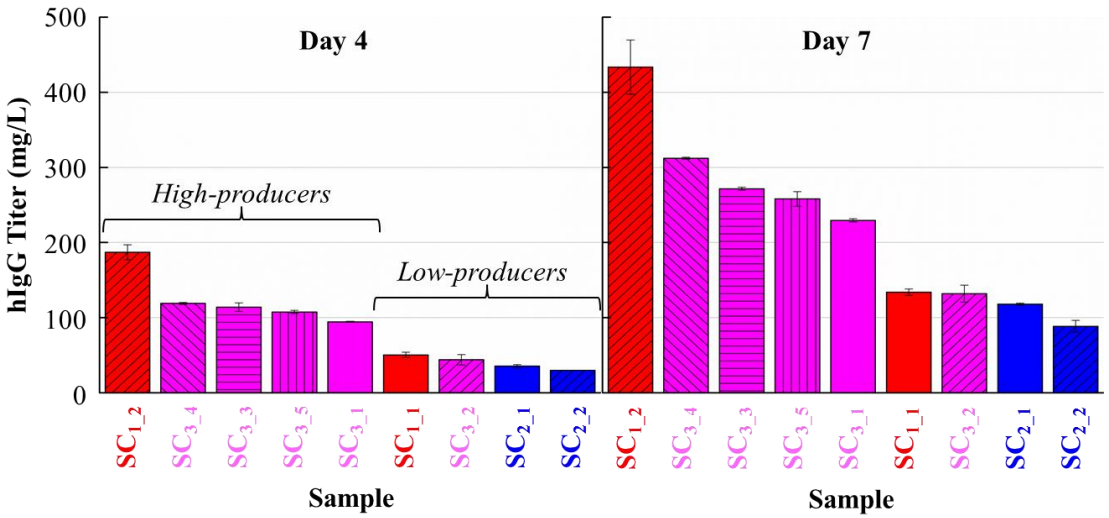


Figure 33: Two classes of sub-clones: high-producers and low-producers with regard to hIgG titer at day 4 (left panel) and day 7 (right panel) of fed-batch cultures.

As presented in Table 6, approximately 1,800 genes were identified as differentially expressed between the five high-producers and the four low-producers at day 4 with a fold-change ≥ 1.5 and a q-value ≤ 10%. At day 7, this number reduced drastically to only 107 genes. Between the two time points, only one gene (Rit2) was commonly up-regulated in high-producers, whereas 41 genes were commonly down-regulated.

Table 6: Number of differentially expressed genes between high and low producing sub-clones at day 4 and day 7 of fed-batch cultures.

Time points / Comparison	Day 4		Day 7	
	Up-regulated	**Down-regulated**	**Up-regulated**	**Down-regulated**
High vs. Low Sub-clones	1153	629	4	103

Functional analysis identified a large number of cellular pathways which were enriched in the high producers at day 4 and day 7 as listed in Table 7. Among the gene sets commonly enriched between these two time points are protein processing (Golgi apparatus and glycan structure biosynthesis), cell cycle, ABC transporter, cytoskeleton function, and signaling pathways (ECM receptor interaction, neuroactive ligand receptor

interaction, HCMV pathway, and TNFR1 pathway). In addition, a few other pathways appeared to be enriched in the high producers only at day 4 such as phenylalanine metabolism, aminoacyl tRNA biosynthesis, DNA replication, cell adhesion, ribosome, MPR and JAK-STAT signaling pathways. At day 7, pyrimidine metabolism, steroids biosynthesis, and FMLP pathway were enriched in the high-producing sub-clones.

Table 7: List of functional classes enriched in high vs. low sub-clones at day 4 and and day 7 of fed-batch cultures.

Gene sets enriched at day 4	**Gene sets enriched at day 7**
Golgi apparatus	Golgi apparatus
Glycan structure biosynthesis	Glycan structure biosynthesis
Cell cycle	Cell cycle
Cytoskeleton function	Cytoskeleton function
ECM receptor interaction	ECM receptor interaction
ABC transporters	ABC transporters
Neuroactive ligand receptor interaction	Neuroactive ligand receptor interaction
HCMV pathway	HCMV pathway
TNFR1 pathway	TNFR1 pathway
Phenylalanine metabolism	Pyrimidine metabolism
Aminoacyl tRNA biosynthesis	Biosynthesis of steroids
MPR pathway	FMLP pathway
DNA replication	
Cell adhesion molecules	
Ribosomal proteins	
JAK-STAT signaling pathway	

5.4.6 Multiple Routes to Hyper-productivity

Given that each clone isolated following selection and each sub-clone isolated following amplification possessed a unique transcriptome signature, it appeared that these cells have developed multiple routes to achieve hyper-productivity. Thus it becomes

critical that a union among the gene sets representing these routes is derived to serve as a universal indication of hyper-productivity traits. To this end, we employed a filtering strategy outlined in Figure 34 to compile a final hyper-productivity gene set.

A set of 349 genes (F) was identified as differentially expressed in high-producers compared to low-producers at day 4 and day 7 of fed-batch cultures using a q-value $\leq$ 10%. This set could subsequently intercept with each of the following three sets to arrive at a final gene list. The first set (S) comprised of 1,996 genes which were differentially expressed in at least one hIgG-producing clone upon selection but not in the control pool. The second set (A) of 988 genes was derived by compiling a union of differentially expressed genes in the three hIgG-producing clones following amplification, also not in the control pool. In these two sets, a slightly stricter q-value cutoff of less than 5% was used to identify differential expression in the clones, whereas genes with q-values passing a commonly used cutoff of 10% were considered unchanged in the control pool. The third set (Q) of 451 genes was compiled from several previous studies of high vs. low-producing NS0 clones (Charaniya et al. 2009; Seth et al. 2007c) and of productivity enhancing culture conditions such as temperature shift and sodium butyrate treatment (Kantardjieff et al. 2010; Yee et al. 2008; Yee et al. 2009).

The intersection between F and S, F and A, and F and Q comprised 58, 13, and 15 genes, respectively. The combined list of these genes is shown in Table 8. The majority of these genes are involved in gene expression and signaling activities (Figure 35).

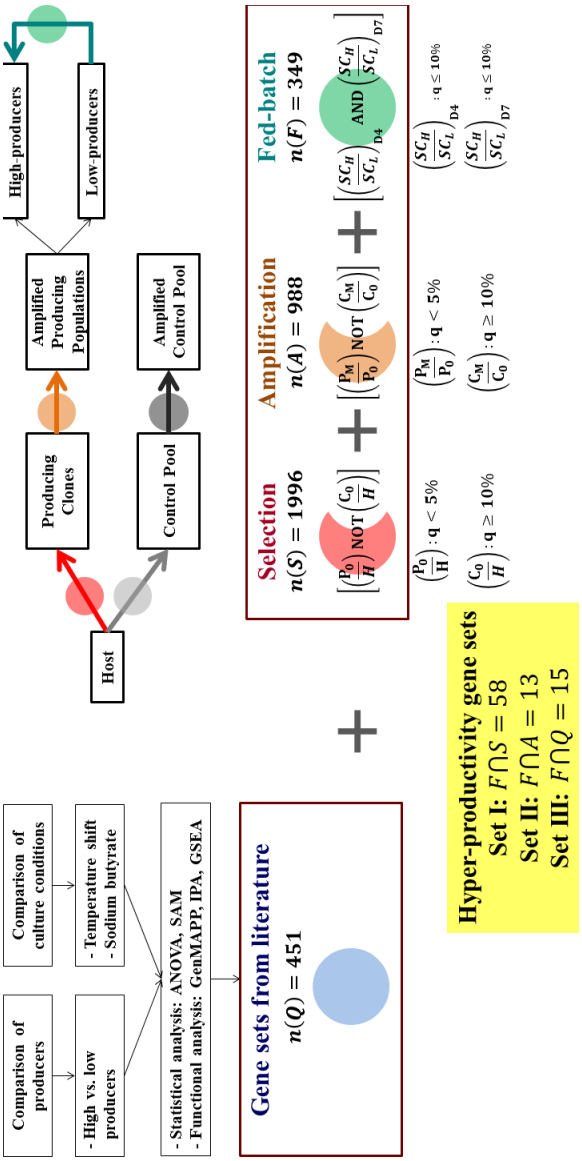


Figure 34: Compilation of hyper-productivity gene sets from multiple sources.

Table 8: List of genes conferring hyper-productivity traits compiled from multiple sources.

Gene symbols	Gene symbols	Gene symbols
2410001C21Rik	Gpatch3	Sar1b
AB370295.1	Gps2	Scamp2
Abcb6	H2afy	Slu7
AC016791.23	Hmgcr	Smc6
AC132224.3	Hsd17b12	Smurf1
Acsl4	Hspd1	Snrpc
Adam9	Ing1	Snx21
AL611927.21	Jmjd4	Stk25
AL824707.7	LOC680531	Stt3b
Arfrp1	Mrps24	Stx7
Arhgap29	Ncapd2	Tdrd3
Atp11b	Nde1	Tfdp1
Ckap4	NM_008350.4	Timp2
CS391352.1	Nt5dc2	Traf6
Cstf2t	Papola	Trim28
Ctbp1	Parp1	Twsg1
Cyp20a1	Pask	Ube2s
D17Wsu104e	Pax3	Ubr1
Dagla	Pcolce2	Ufm1
Dnpep	Pin4	Utp11l
Dpm1	Prpf31	Wtap
E2f6	Ptpn14	Xrcc5
Fam3a	Ptrf	Ythdf1
FI847326.1	Pxk	Zfand5
Fn1	Rac1	Zfp128
FQ075827.1	Rpe	Zfp346
GH511842.1	Rps20	Zfp600
GH524702.1		

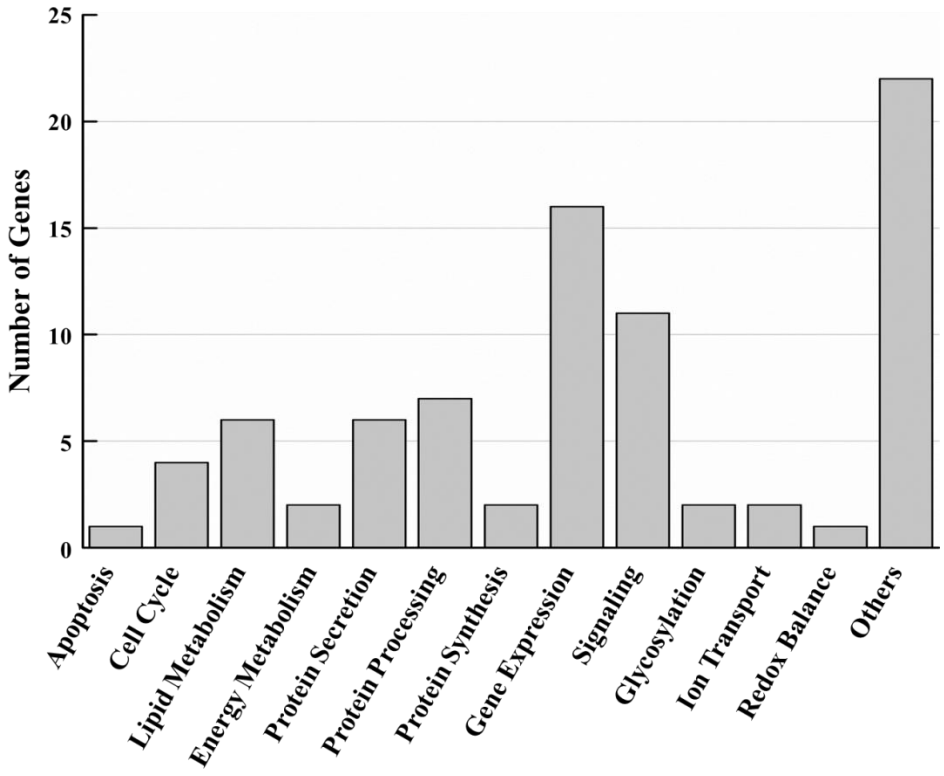


Figure 35: Distribution of functional enrichment in the hyper-productivity gene set.

5.5 DISCUSSION

The process of initial selection followed by amplification has been used widely in the past few decades for generating recombinant protein producing cell lines. However, very limited information is currently available regarding the transcriptional changes incurred by this transformation process. To our knowledge, this is the first study that employed a genome-scale transcriptome approach to investigate these changes. Hundreds of genes were identified as differentially expressed upon selection and amplification, most of which are related to signaling, cell cycle, DNA replication, and amino acid metabolism.

A moderate variability was observed among the three hIgG-producing clones following selection, which harbored the plasmid at random locations in the host genome. Following amplification, this difference was expanded significantly. This is likely a

result of the "position effect", in which the integration site of the plasmid would significantly affect the expression of the transgenes (Wilson et al. 1990). Integration of the plasmid into a heterochromatin region can lead to little or no expression, whereas integration into a euchromatin region often allows active expression of the transgenes. Another contributing factor could be a difference in the copy number of plasmid integrated into the genome. In a different study, it was shown that the range of plasmid copy number can be between 20 and 400 in six randomly chosen clones (Wurm and Petropoulos 1994). Furthermore, from hierarchical clustering results, different samples appeared to cluster amongst clones rather than across time. This result suggests that each clone may possess a unique transcription signature which could dictate the cells' propensity to amplify and which was well retained even after amplification.

Given the assumed roles of selection and amplification, it is surprising to find that selection conferred more drastic transcriptional changes of the transgenes than did amplification. Except for DHFR, which was induced at comparable levels, the other three transgenes (namely HPT, hIgG light chain, hIgG heavy chain) experienced a more than one-thousand-fold increase upon selection relative to amplification. This result indicated that selection was not simply selecting for cells with stable integration of the selection marker. Indeed, during the selection process, cells might have been forced to expand other cellular functions to cope with the selection pressure such as cell cycle, DNA replication, and amino acid metabolism. A comparison between hIgG-producing clones and the control pool further revealed several aspects of the cellular machinery which are potentially required for secreting hIgG at high levels. Such cellular functions encompass glutathione metabolism, mRNA processing, and several signaling pathways such as mTOR, EDG1, and PDGF. Not only does amplification appear to incur modest changes in the expression levels of the transgenes, it also induces minor changes at a global level with respect to selection. Almost identical cellular pathways were shown to be enriched following amplification. Thus, amplification seems to play a role in further reinforcement of transcriptional changes which have already occurred during selection rather than giving rise to new and unique changes.

A comparison between high and low-producing sub-clones derived from amplified populations in fed-batch cultures uncovered additional cellular pathways necessary for secreting hIgG efficiently. In high-producing sub-clones, Golgi apparatus, glycan structure biosynthesis, cell cycle, cytoskeleton function, ECM receptor interaction, ABC transporter, and several signaling pathways were significantly enriched compared to low-producing sub-clones. Among these pathways, protein processing, cell cycle, and cytoskeleton-related elements were also identified in a previous transcriptome study of seven high- and four low-producing clones of NS0 cells (Charaniya et al. 2009). In another study, combined transcriptome and proteome analysis of CHO cells revealed enrichment of secretory pathways, including Golgi apparatus, cytoskeleton protein binding, and small GTPase-mediated signal transduction under temperature shift and sodium butyrate treatment (Kantardjieff et al. 2010).

Given that there is such a high degree of dissimilarity among different clones, each of which has altered a large number of genes and functional classes to become high-producers, it becomes evident that hyper-productivity can be achieved via multiple routes. Indeed, this concept has been articulated previously (Seth et al. 2007a). As shown in Figure 36, hyper-productivity can be considered a complex trait that requires a collection of many positive characteristics encompassing a wide range of functions, including energy metabolism, protein secretion, redox balance, and growth/death control.

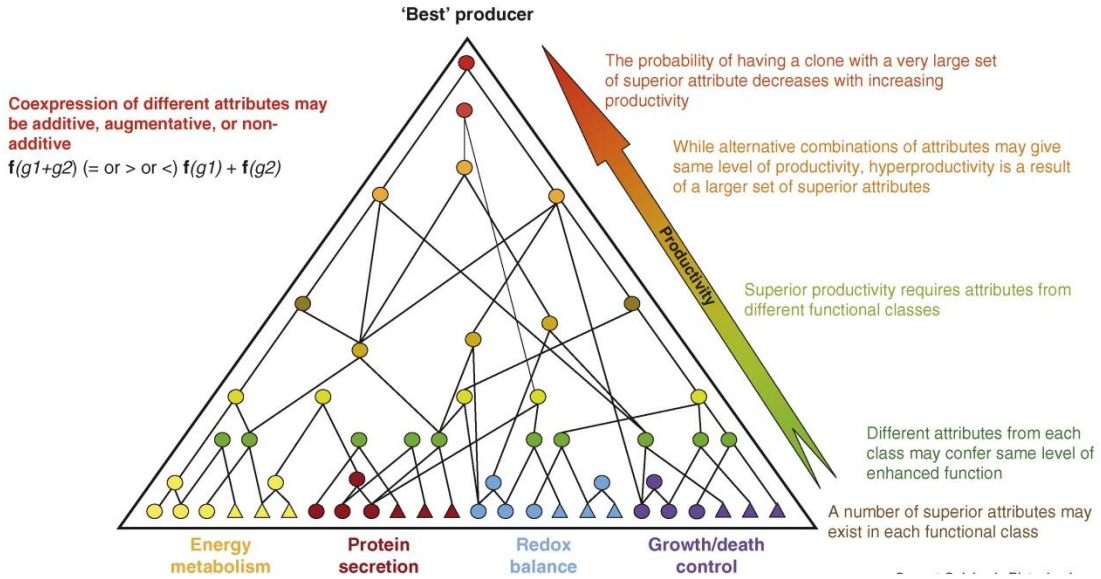


Figure 36: Multiple routes to hyper-productivity (Seth et al. 2007a).

The surprisingly modest role of amplification relative to to selection observed in this study suggests that selection itself may suffice to generate a wide range of clones to screen for high-producers. Furthermore, it would be interesting to explore the long-term stability of such high product titers achieved upon cell line development. It has been shown that transgene mRNA levels can potentially decline over time in culture, possibly due to repeat-induced silencing of the amplified product gene (Chusainow et al. 2009). Thus more mechanistic insights would be required to derive a conclusive guide for cell line development process. As more genomic tools for CHO cells are becoming available, we can anticipate an abundance of such studies characterizing different aspects of hyper-productivity traits in the years to come.

6 ENGINEERING GENE EXPRESSION DYNAMICS OF MAMMALIAN CELLS IN CULTURE

6.1 SUMMARY

Cultured mammalian cells are the major hosts for the production of recombinant protein therapeutics. Genetic engineering has been used to transform these cells from non-producers to hyper-producers, which harbor superior phenotypic traits related to metabolism, protein secretion, and growth control. Many of these efforts have been performed using strong and constitutive expression systems. However, cellular needs are dynamic, responding to various environmental perturbations as well as cellular processes, rather than being static. Inducible systems enable time dynamics through external manipulation of inductive conditions. Ideally, a transgene's expression should be synchronous with the host cell's own rhythm. Furthermore, depending on the cellular process to be manipulated and the transgene's function, the expression level should be controlled at the correct modulating level.

To that end, we surveyed the transcriptome profiles obtained from time-course experiments spanning various stages of cell growth and encompassing different fed-batch culture conditions. Clustering of gene expression patterns revealed thousands of genes with varying dynamics across a wide intensity range. The promoters of these genes can potentially provide a more flexible and dynamic means to control the expression of the transgene.

In this study, we focused on several genes which exhibit low expression levels in the mid-exponential phase and significantly higher expression levels in the late-stationary phase. We demonstrated that, under the control of these dynamic promoters, a blasticidin resistance (BSD) gene and an enhanced green fluorescent protein (EGFP) gene can be expressed in concert with cell growth. This approach illustrates a novel concept in metabolic engineering which can potentially be used to achieve dynamic control of cellular behaviors for enhanced process characteristics.

6.2 INTRODUCTION

Recombinant mammalian cells are used extensively for the production of therapeutic proteins, accounting for tens of billions per annum of product value and benefiting tens of thousands of patients (Aggarwal 2011). Their rise to such prominence largely has been attributed to genetic manipulations which enable cells to produce high levels of recombinant proteins. Substantial efforts in genetic engineering have focused on introducing metabolic, secretory, and growth control genes. Significant progress has been reported, including down-regulation of lactate dehydrogenase A (LDH-A) and pyruvate dehydrogenase kinases (PDHK) to reduce lactate production (Zhou et al. 2011), functional disruption of an α-1,6-fucosyltransferase (Fut8) to produce nonfucosylated hIgG1 (Malphettes et al. 2010), and over-expression of anti-apoptotic genes E1B19K and Aven to delay the onset of apoptosis (Figueroa et al. 2007).

Typically, genetic engineering of these target genes in cultured mammalian cells involves the use of either constitutive or inducible promoters. Constitutive promoters such as those derived from Cytomegalovirus (CMV), Simian virus 40 (SV40), or the elongation factor 1 α (EF1α) are virtually always active regardless of environmental conditions. Such strong and constitutive expression of the transgenes may not be suitable in several cases wherein the transgenes have negative impacts on cellular behaviors. Inducible promoters, alternatively, are activated upon addition of the corresponding inducers such as tetracycline, isopropyl-b-D-thio-galactoside (IPTG), or hormones; thus, inducible systems allow more flexibility in modulating gene expression. However, the presence of these inducers could potentially trigger undesirable cellular responses. Furthermore, gene expression is a dynamic process by nature, which is constantly adjusting to cope with various perturbations. Consequently, endogenous promoters with intrinsic dynamic activities represent attractive tools to control the expression dynamics of the transgenes.

The vast increase of time-course transcriptome data in recent years, mainly through microarray and deep sequencing approaches, presents unprecedented opportunities to discover genes with dynamic expression profiles. Indeed, analysis of

time-course deep sequencing (deepCAGE) data revealed the time-dependent dynamics of the 30 most significant regulatory motifs in a human cell line (Suzuki et al. 2009). Such variety of time dynamics was also observed in an hIgG-producing CHO cell line undergoing temperature shift and sodium butyrate treatment (Kantardjieff et al. 2010). In these studies, the major expression trends of hundreds of regulatory motifs or genes were distinguished from the background noise either by model fitting or principal component analysis (PCA). These trends subsequently could be classified into several groups by clustering based on their similarity. The promoters of these genes, once isolated, would serve as potential tools for dynamic control of gene expression.

Recent advent of genome sequences for the most important industrial cell lines, Chinese Hamster Ovary (CHO) and Baby Hamster Kidney (BHK) cells, has permitted further avenues for dynamic cell engineering (Broad Institute 2011; Jacob et al. ; Xu et al. 2011). Although not as completed as the human or mouse genomes, it would be expected that regulatory elements in these cells can be identified and isolated. Pioneering work by the ENCODE and the FANTOM consortiums have revealed that transcriptional regulation in mammalian cells is much more complex than previously known. Gene expression is regulated by an intertwined network which include histone modifications; dispersed enhancers, silencers, and insulators; and alternative promoter usage, splicing, and polyadenylation (The ENCODE Consortium 2007; The FANTOM Consortium 2005). The concept of a promoter region thus becomes fuzzy, as it may encompass thousands of base pairs (bp) and contain a large number of regulatory elements. In the scope of this study, we focused on the commonly known core and proximal promoter regions, which are the most well-defined regions that can be routinely isolated. The core promoter often resides within a few hundred bp of the minimum portion required for transcription initiation, which involves binding sites of the RNA polymerase and general transcription factors (TFs) (Butler 2002). The proximal promoter region, typically extended up to a few thousand bp, contains the core promoter and most of the specific TF binding sites.

In this study, we surveyed historical time-course microarray data from multiple fed-batch cultures to identify genes with time dynamic expression trends. The promoter

of several candidate genes were isolated from Chinese hamster liver genomic DNA, and used to drive the expression of a fusion transgene following the expected dynamic manner. Thus this study represents the first proof-of-concept in dynamic cell engineering.

6.3 MATERIALS AND METHODS

6.3.1 Time-series Transcriptome Data Processing

Transcriptome data obtained from 72 samples from twelve fed-batch cultures of a recombinant CHO cell line producing IgG were used in this study. RNA samples of six time points in each culture were obtained and assayed with Affymetrix microarray. Raw data (.CEL) files were processed using the GeneData Expressionist Refiner module (GeneData, San Francisco, CA). The Microrray Analysis Suite Statistical Algorithm (MAS 5.0) was used to assess the overall quality and to condense the intensities of all probes within each probe set to a single value. The average intensity value of all probe sets within each array was linearly scaled to 500. Quantile normalization was further applied to align the intensity distributions of all arrays (Bolstad et al. 2003). Probe sets with a minimum detection *p*-value greater than 0.04 or a maximum intensity value less than 70 in all samples were called absent and removed from further analysis. Since we were only interested in transcripts which are dynamically expressed, "invariant" probe sets with a coefficient of variation (CV) across time less than 20% in all samples were also excluded. If a gene is represented by multiple probe sets, the one with highest CV was selected.

6.3.2 Time Profile Patterns of Transcriptome Data

Log_2-mean centered intensity values for each gene were arranged into a 12 x 6 matrix, representing the twelve fed-batch cultures, each with six time points. Principal component analysis (Alter et al. 2000) was performed on each matrix (Spotfire, Cambridge, MA). Genes for which the first principal component (PC1) captured more than 80% of the total variance were retained for further analysis. These genes were grouped into clusters with similar expression time profiles using k-means clustering (Everitt 1974a) (k = 6) with their first principal components as the input. Genes within

each of the six clusters were furthered categorized into high, mid, and low intensity ranges with the maximum intensity across samples (I_{max}) greater than 10000, between 10000 and 5000, and less than 5000, respectively. Finally, functional classes enriched amongst each of the six clusters (i.e., enrichment *p*-value less than 0.05) were identified using gene ontology enrichment analysis implemented in GenMAPP's MAPPFinder (Doniger et al. 2003).

6.3.3 Fed-batch Cultures

The transcript dynamics of the candidate promoter was verified in a second recombinant CHO cell line prior to promoter isolation. The same cell line was used as the host cell line for expressing the promoter-reporter construct. The cultures were initiated by inoculating cells at 2.5 x 10^5 cells/mL into 20 mL of DMEM-F12 medium in each 125 mL shaker flask (Thermo Scientific, Rochester, NY) and incubated at 37^{o}C on a shaker rotating at 130 rpm, in 5% CO_2 environment in a humidified incubator. One mL of a ten-fold concentrated feed medium (10 X) was added daily starting from day 2. Cell concentration and viability were determined by counting with a hemacytometer using trypan blue staining. One million cells were withdrawn each day for RNA extraction starting from day 2.

6.3.4 Quantitative Real Time PCR (qRT-PCR)

Total RNA was isolated using the RNeasy Mini kit (Qiagen, Valencia, CA) according to the manufacturer's protocol with on-column DNase I treatment for removal of genomic DNA. Reverse transcription was performed with 2.5 μg of total RNA using Superscript III Reverse Transcriptase (Invitrogen, Carlsbad, CA) with 1 mM oligo dT primers in a total volume of 50 μL. A no reverse transcriptase control was performed in parallel to assess genomic DNA contamination.

qRT-PCR primers for several dynamic genes, listed in Table 9, were designed using the Primer3 website (http://frodo.wi.mit.edu/). The mRNA levels of these genes were quantified using the Brilliant II SYBR Green qPCR Master Mix (Agilent, Santa Clara, CA). Each 12.5 μL reaction mixture contained 6.25 μL of the master mix, 50 ng of the cDNA template, and 0.2 μM of each primer (forward and reverse). qRT-PCR was

performed using the Stratagene Mx3000P instrument (Agilent, Santa Clara, CA). The thermal cycling profile was set as follows: initial denaturation at 95°C for 10 min, followed by 40 cycles of 95°C for 10 sec, 57°C for 1 min, and 72°C for 30 sec. The dissociation curves of the PCR products were generated by ramping from 57°C to 95°C after a denaturation step at 95°C for 1 min and an annealing step at 57°C for 30 sec. All cDNA samples were run in triplicates along with a no reverse transcriptase control and a no template control. β-actin was used as a reference for comparison across samples.

Table 9: List of qRT-PCR primers used for quantification mRNA levels of candidae genes, β-actin, BSD, and EGFP.

Gene symbol	**Ensembl ID of mouse homolog**	**Left primer**	**Right primer**
Mmp12	ENSMUSG00000049723	CAGCCATCTTTGACCCATCT	GAGCCTTTTGGTGACACGAT
Sppr1a	ENSMUSG00000050359	AACCAAGGATCCCTGCAAC	CATGGCTCAGGAACAACTGG
Gsta3	ENSMUSG00000025934	GCCAAGATCAAGGACAAAGC	GTCCAGCTCTTCCACATGGT
Hmox1	ENSMUSG00000005413	CCTAAAGCGGACAGAACCAG	ACCTGGCCCTTCTGAAAGTT
Psap	ENSMUSG00000004207	CTGTCCAAGACCCGAAGGTA	TTTCAGCAAGTTCCCAGCTT
Clu	ENSMUSG00000022037	AAATTCAAAATGCCGTCCAG	CCATCATGGTCTCATTGCAC
Ces1f	ENSMUSG00000031725	TTGGAGAGTCAGCAGGAGGT	CAGATGTGGTGGTTTTGCAC
Serpinf1	ENSMUSG00000000753	TCAAGGTCCCGGTAAACAAG	GGTGCTATGGATGTCCGAGT
Txnip	ENSMUSG00000038393	CAGTGCAAACAGACCTTGGA	AAGGAAAGCCTTCACCCAGT
Sqstm1	ENSMUSG00000015837	CTACACAGGGAGCACAGCAA	ATCCCCTGCTCTAGGAGGAC
Nucb2	ENSMUSG00000030659	CTCAGTGGCCTCATCTGTGA	TCCTGGTGGGTCTATCCTTG
Trp53inp1	ENSMUSG00000028211	TTTCCAATTCCCATGCAGAT	AGGACGGAGCAAAATAGCAA
Stau2	ENSMUSG00000025920	GAGAGCCTGCCATCTACAGG	CATTGTGTCTGGCAGCTTGT
Lcp1	ENSMUSG00000021998	ATGCCCTGATCATCTTCCAG	TGATTCTTCCCCAGATCCAC
CD36	ENSMUSG00000002944	CCATCTACGCTGTGTTTGGA	TGTTTGCATTTGCTGATGTCT
β–actin	ENSMUSG00000029580	GTCGTACCACTGGCATTGTG	AGGGCAACATAGCACAGCTT
BSD	N.A.	ATGCAGATCGAGAAGCACCT	ATCAACAGCATCCCCATCTC
EGFP	N.A.	ACGTAAACGGCCACAAGTTC	AAGTCGTGCTGCTTCATGTG

6.3.5 Isolation of Chinese Hamster Promoter

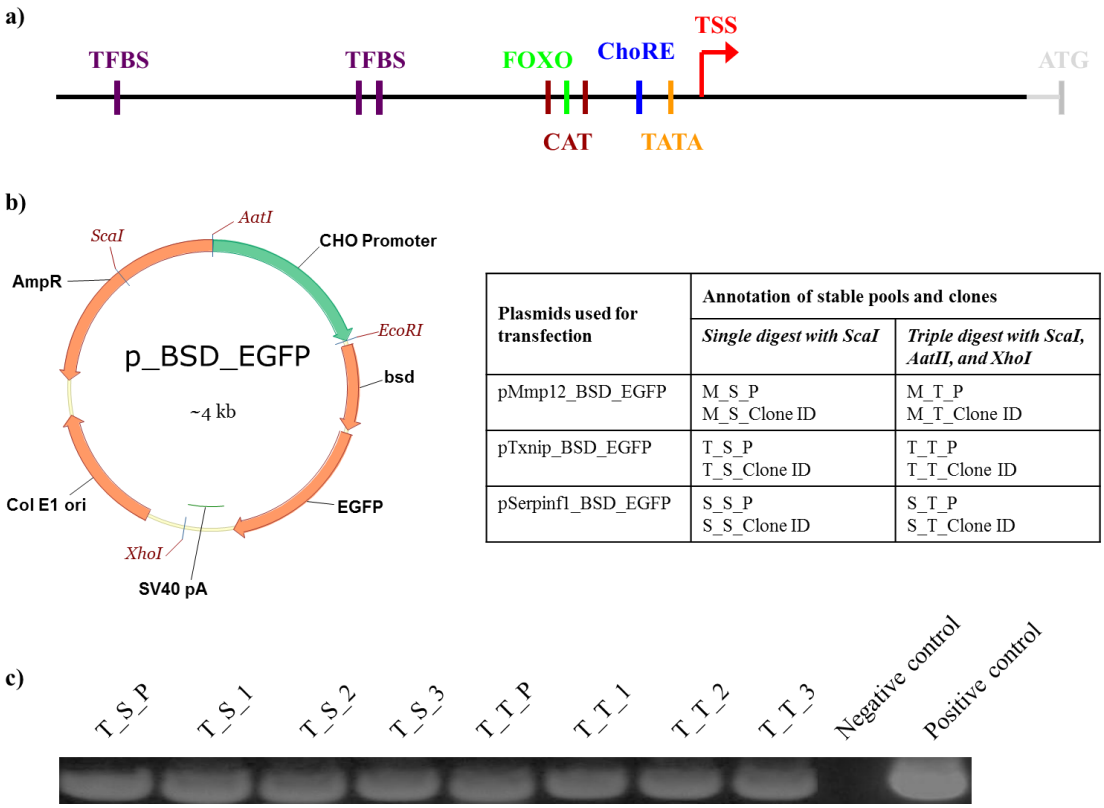
A fragment of approximately 1000 bp containing part of the putative 5' UTR and the upstream region of each of the following genes, Mmp12, Txnip, and Serpinf1, was isolated. Genomic DNA isolated from Chinese hamster liver with the DNeasy Blood and Tissue kit (Qiagen, Valencia, CA) was used as the PCR template. The forward and reverse primers included restriction sites of AatII and EcoRI, respectively, to facilitate cloning of the PCR product into the expression vector (Table 10).

Table 10: List of PCR primers used for isolating promoters of Mmp12, Txnip, and Serpinf1 from Chinese hamster (CH) liver genomic DNA.

Letters in regular format represent DNA sequences complement to the template. Underlined letters represent restriction sites used for cloning. Italic letters represent random sequences needed to make the restriction sites "internal".

Gene symbol	**Mmp12**	**Txnip**	**Serpinf1**
Left primer	*TCAG*GACGTCGGTGG GAGTGTGTGTTCCTT	*CGTA*GACGTCCCAATG CTGAAGAACCCTTG	*TCGA*GACGTCCACCTC ACTGGCCACTTTTT
Right primer	*TCAG*GAATTCTGTGAC CAGATCTCTCAGCAG	*TCAG*GAATTCTCAGCG GGTTCCAGATAAAC	*TCGA*GAATTCCTCTAG CAAGCAGGGGAGTG
Coordinate in Chinese hamster scaffolds (Jacob et al.)	1274477..1275388, scaffold 8	1623404..1624318, scaffold 122	60652..61866, scaffold 942
Coordinate in CHO-K1 scaffolds (Xu et al. 2011)	709505..710405, scaffold 370	846522..845735, scaffold 297	56467..57667, scaffold 16294

A 50 μL PCR reaction mixture contained 1 unit of the Phusion High Fidelity DNA Polymerase (Finnzymes, Vantaa, Finland), 100 ng of the liver genomic DNA template, 0.5 μM of each primer, and 200 μM of dNTPs (Invitrogen, Carlsbad, CA). The reaction was performed under the following conditions: initial denaturation at 98^{o}C for 30 sec, 30 cycles of 98^{o}C for 10 sec, 57^{o}C for 30 sec, and 72^{o}C for 30 sec prior to final extension of 72^{o}C for 10 min. The products were separated on a 0.8% agarose gel. The expected band of approximately 1000 bp was excised and purified using the Zymoclean Gel DNA Recovery kit (Zymo Research, Irvine, CA). A representation of the Txnip promoter fragment is shown in Figure 37a.



Plasmids used for transfection	Annotation of stable pools and clones	
	Single digest with ScaI	*Triple digest with ScaI, AatII, and XhoI*
pMmp12_BSD_EGFP	M_S_P M_S_Clone ID	M_T_P M_T_Clone ID
pTxnip_BSD_EGFP	T_S_P T_S_Clone ID	T_T_P T_T_Clone ID
pSerpinf1_BSD_EGFP	S_S_P S_S_Clone ID	S_T_P S_T_Clone ID

Figure 37: Cloning of promoter fragments from Chinese hamster liver genomic DNA.

a) Representation of Txnip promoter fragment (black horizontal line) with approximate locations of putative transcription start site (TSS); general transcription factor binding sites (TFBSs) such as TATA box and CAT box; and other specific transcription factor binding sites such as FOXO binding site and carbohydrate response element (ChoRE). The start codon (ATG) was shown for reference.

b) Plasmid map of p_BSD_EGFP. The isolated promoter was used to drive the expression of a fusion gene composed of a blasticidin resistance marker (BSD) and an enhanced fluorescent protein (EGFP) gene. Each plasmid was linearized with restriction enzymes prior to transfection in two different ways: single digest with ScaI, or triple digest with ScaI, AatII, and XhoI. Thus the resulting stable pools and clones are labeled as A_B_C, where A stands for the promoter (M for pMmp12, T for pTxnip, and S for pSerpinf1), B stands for the type of restriction digest (S for single digest and T for triple digest), and C stands for the pool (P) or the clone ID (numeric value).

c) Integration of plasmid into the host cells' genome was verified using gel electrophoresis of the PCR product of genomic DNA isolated from transfected cells. The left and right primers used for PCR flank the BSD and the EGFP genes, respectively.

6.3.6 Construction of Expression Vector

A fusion gene of a blasticidin resistance marker (BSD) and an enhanced green fluorescence protein (EGFP) was excised from the CSII_BSD_EGFP plasmid (courtesy of Dr. Nikunj Somia) using NotI and EcoRI. The pTRE-Tight plasmid (Clontech, Mountain View, CA) was also digested using the same combination of restriction

enzymes. Each restriction digest mix contained 1 μg of plasmid, 10 units of NotI and 20 units of EcoRI (New England BioLabs, Ipswich, MA). The digest was performed at 37^{o}C for 2 hr following by heat inactivation at 65^{o}C for 20 min. The digested products were separated in a 0.8% agarose gel, and purified using the Zymoclean Gel DNA Recovery kit (Zymo Research, Irvine, CA).

The fusion gene was inserted into the multiple cloning site of pTRE_Tight to generate a new vector, pTet_BSD_EGFP. About 50 ng of the BSD_EGFP insert and 50 ng of the pTRE_Tight backbone were ligated overnight at 16^{o}C in a 20 μL ligation mix with 0.5 unit of T4 DNA Ligase (Invitrogen, Carlsbad, CA). A negative control was performed in parallel without the insert to assess the extent of self-ligation of the backbone. Two μL of this ligation mix was used to transform chemically competent *E.coli* cells OneShot TOP10 (Invitrogen, Carlsbad, CA) following the manufacturer's protocol. Twelve single colonies were randomly selected and subjected to PCR amplification using two primers: EGFP_N (CGTCGCCGTCCAGCTCGACCAG) and BSD_R1 (ATCAACAGCATCCCCATCTC) to identify colonies with the correct insert direction. Two of them were chosen to inoculate overnight cultures in 5 mL of LB medium. The amplified plasmids were purified from 4.5 mL of the overnight culture using the Qiaprep Spin Miniprep kit (Qiagen, Valencia, CA). The pTet_BSD_EGFP plasmid sequence in each colony was verified using Sanger sequencing with the EGFP-N primer.

The tetracycline inducible promoter in this plasmid was subsequently replaced by each of the three Chinese hamster promoters isolated above via the restriction sites of AatII and EcoRI in a similar cloning process. These plasmids were named after the corresponding promoters: pMmp12_BSD_EGFP, pTxnip_BSD_EGFP, and pSerpinf1_BSD_EGFP (Figure 37b).

6.3.7 Generation and Characterization of Stable Pools and Clones

The plasmids constructed above were linearized prior to transfection into the host cells. Two types of restriction digest were used: one with ScaI (single digest), which cuts inside the ampicillin resistance gene; and the other with ScaI, AatII, and XhoI (triple

digest) to remove all bacterial components. The products of the triple digest were separated on a 0.8% agarose gel, and the expected band was excised and purified using the Zymoclean Gel DNA Recovery kit (Zymo Research, Irvine, CA).

CHO cells were seeded in duplicated wells at 5 x 10^5 cells/mL in 2 mL of DMEM-F12 medium in a 6-well plate 24 hr prior to transfection. A DNA – cation lipid complex containing 4 μg of linearized plasmid, 10 μL of Lipofectamine 2000 (Invitrogen, Carlsbad, CA), and 500 μL of Opti-MEM (Invitrogen, Carlsbad, CA) was generated according to the manufacturer's protocol. After an incubation of 20 min at room temperature, the complex was added drop by drop into the cells. The transfected cells were centrifuged at 700 rpm for 5 min at room temperature 6 hr after transfection and re-suspended in 2 mL of fresh DMEM-F12 medium.

Twenty-four hours following transfection, the cells were diluted in 96-well plates at 1000 cells/well in 0.2 mL of the selective medium comprising DMEM-F12 supplemented with 5 μg/mL Blasticidin S (InvivoGen, San Diego, CA), 5% (v/v) fetal bovine serum (Atlas Biologicals, Fort Collins, CO), and 20% (v/v) conditioned medium. The conditioned medium was collected during the mid-exponential growth phase of the non-transfected cells in a batch culture in DMEM-F12, and was filtered through a 0.45 μm pore size filter (Fisher Scientific, Pittsburgh, PA). In parallel, the pools were also diluted at 5 x 10^5 cells/mL in the same selective medium in 6-well plates and passaged every three days. The plates were incubated for approximately two weeks in a 37^{o}C, 5% CO_2 environment. Single clones in 96-well plates were picked and transferred to a new well containing 150 μL of fresh selective medium. Upon reaching high cell density, the total volume of 200 μL was expanded gradually to 1 mL in 24-well plates, 4 mL in 6-well plates, 8 mL in T-25 flasks, and 25 mL in T-75 flasks for cryopreservation and subsequent characterization in the selective medium.

Stable clones and pools were characterized for the mRNA levels of the transgenes (BSD and EGFP) and the endogenous genes across different stages of fed-batch cultures via qRT-PCR. A simple flowchart describing the process of generating and characterizing stable pools and clones is presented in Figure 38.

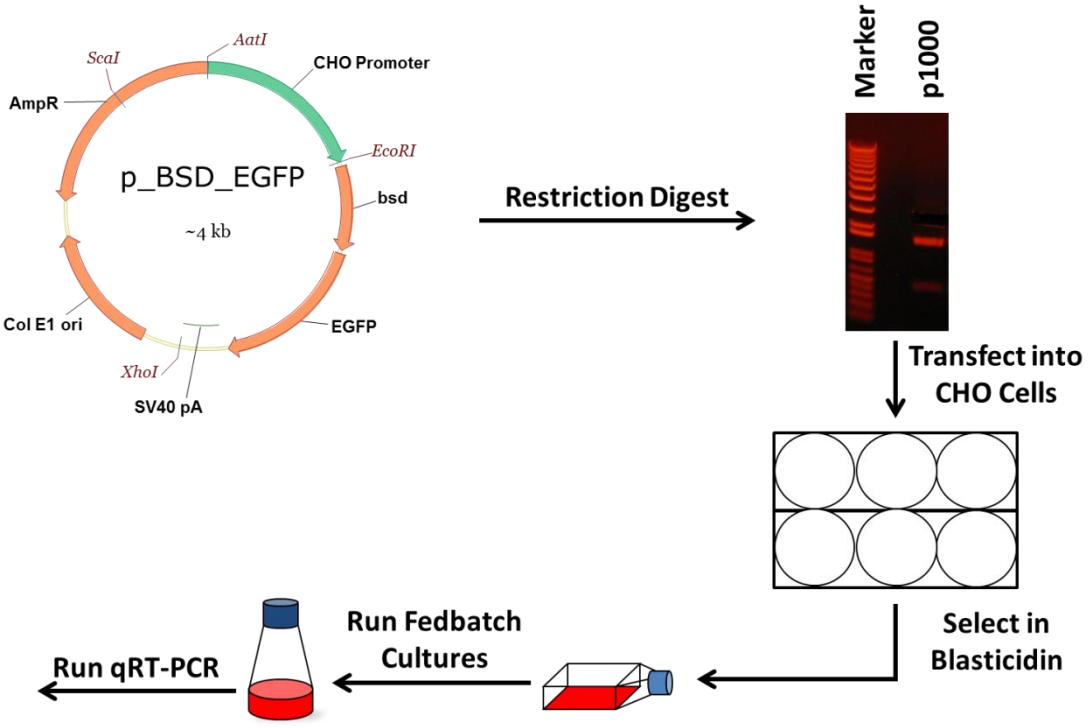


Figure 38: Flowchart for clone generation and characterization process.

The plasmids were linearized before transfection either by single digest with ScaI or triple digest with ScaI, AatII, and XhoI. The digested products were separated and purified prior to transfection. Transfected cells were selected for two weeks in blasticidin. Stable pools and single clones were isolated and used to run fed-batch cultures. Samples were collected over time to characterize mRNA levels of the transgenes (BSD, EGFP and the endogenous genes.

6.4 RESULTS

6.4.1 Identification of Genes with Dynamic Expression Profiles

Historical time-series microarray data obtained from multiple fed-batch culture conditions were used to identify genes with reproducible dynamic behaviors in a typical fed-batch culture. Among the 23020 probe sets available on the Affymetrix microarray, 10256 represented genes which were expressed at varying levels across time as evaluated by a coefficient of variation (CV) of greater than 0.2. Principal component analysis was performed on each of these genes (Figure 39a). Over 3000 genes with similar expression trends across culture conditions were grouped into six clusters using k-means clustering.

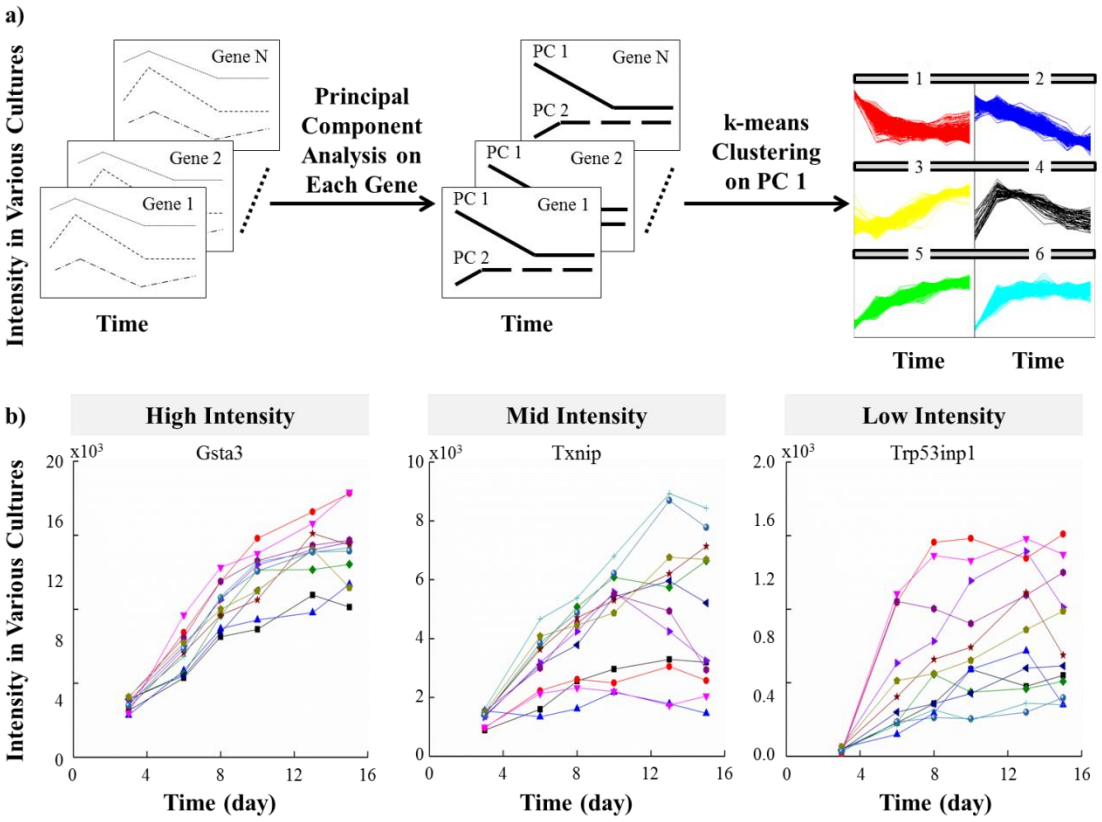


Figure 39: Identification of genes with dynamic expression profiles using historical time-course microarray data.

- **a) Principal component analysis (PCA) was performed on each of the genes with a coefficient of variance (CV) across time greater than 20%. For genes with PC 1 captured more than 80% of the total variance, k-means clustering (k=6 in this case) was used to identify clusters of genes with similar major expression trends. Genes in each cluster follow a distinct expression profile and span a wide range of intensity.**
- **b) Expression profiles of three genes in the 5th cluster in the 12 historical fed-batch cultures are shown, each in a different color. Each gene represents one of the three intensity categories: high, mid, and low with the maximum intensity greater than 10000, between 10000 and 5000, and lower than 5000, respectively.**

Of those, clusters 3, 5, and 6, comprising 1581 genes with low expression level at mid-exponential phase and significantly higher expression level at late-stationary phase, were of special interest. Members of each cluster were further categorized into sub-clusters based on their maximum intensity values (I_{max}): high intensity with I_{max} greater than 10000, mid intensity with I_{max} between 10000 and 5000, and low intensity with I_{max} less than 5000. Examples of three genes representing these three sub-clusters are shown in Figure 39b. Despite spanning a wide intensity range, all three genes displayed similar

expression trends across culture conditions, demonstrating the effectiveness of our time-series transcriptome data mining scheme.

A number of functional classes were significantly enriched in these clusters as determined using gene ontology enrichment analysis (Table 11). Of note, the activity of cell cycle, mRNA processing, ribosome biogenesis, and lipid biosynthesis appeared to decrease gradually over time; whereas apoptosis, protein secretion, redox balance, and fatty acid metabolism displayed increasing trends as the cultures progressed from the mid-exponential phase to the late-stationary phase.

Table 11: Enriched gene ontology classes in the six clusters of dynamic genes.

k-means cluster	Number of genes	Enriched gene ontology classes
1	683	DNA binding, cytoskeleton, cell cycle, mRNA transport and processing, lipid biosynthesis
2	697	Ribosome biogenesis, cell cycle, amine metabolism, mRNA processing
3	429	Actin cytoskeleton, vacuole, vesicle-mediated transport, fatty acid metabolism, sphingolipid metabolism
4	54	No significantly enriched classes
5	757	Glutathione transferase activity, cation transporter, positive regulation of apoptosis
6	395	Apoptosis

6.4.2 Verification of Dynamic Expression Profiles

Among the 1581 genes in clusters 3, 5, and 6, only 901 were annotated with high confidence levels by similarity in nucleotide sequence with mouse, human, or rat homologs. We further selected 15 genes based on the availability of full-length mRNA sequence in Chinese hamster; availability of genomic DNA sequence surrounding the coding region and approximately 1000 bp upstream of the putative transcription start site in Chinese hamster; and validity of protein function of homologs in other species (Table 10). As shown in Figure 40a, in a representative historical culture, these 15 genes displayed similar increasing trends over time with diverse intensity ranges. Thus they

represented ideal candidates for which promoters could be used to drive the expression of transgenes following the desired dynamic trend above.

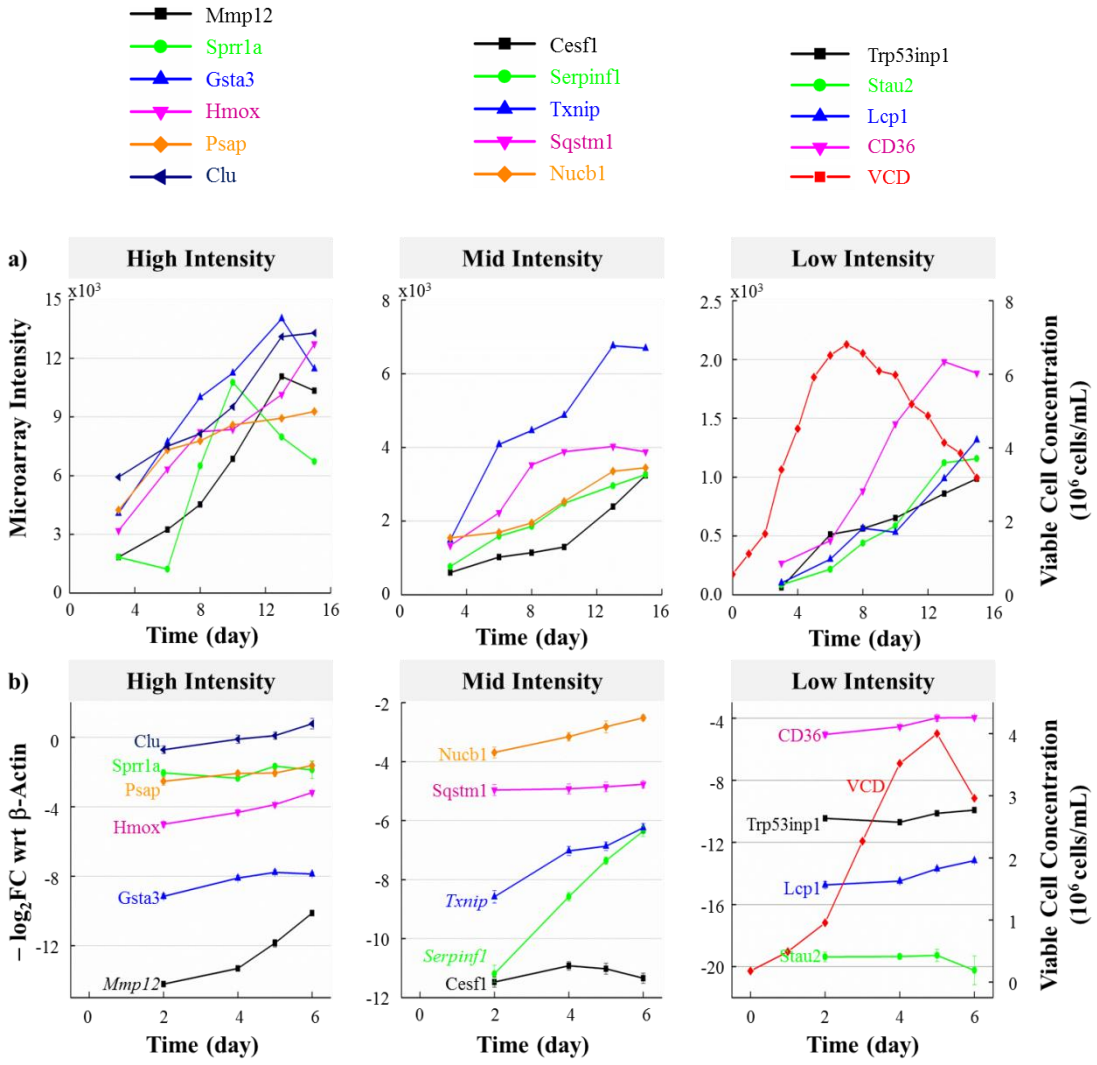


Figure 40: Expression time profiles of the 15 final candidate genes in three categories: high, mid, and low intensity.

- **a) Intensity time profiles of these genes in a representative historical fed-batch culture of an industrial CHO cell line were shown, each in a different color. The viable cell concentration in this culture was also shown for reference.**
- **b) Difference in qRT-PCR cycle number (Ct value), or – log2 fold-change, of each gene with respect to β-actin in a typical fed-batch culture of a second CHO cell line (host). The same color code is used for each gene. The viable cell concentration of this culture was also shown for reference.**

In order to check whether these 15 candidate genes are expressed in the second CHO cell line following the dynamic trends observed in historical data, qRT-PCR

analysis was performed using samples from a typical fed-batch culture of these cells. Surprisingly, only three of them exhibited significant increases in their expression over time as expected, namely Matrix metallopeptidase 12 (Mmp12), Thioredoxin interacting protein (Txnip), and Serine (or cysteine) peptidase inhibitor, clade F, member 1 (Serpinf1). As shown in Figure 40b, when β-actin was used as a baseline for comparison across samples, the expression levels of three genes increased 4- to 28-fold when the cells progressed from the mid-exponential phase to the late-stationary phase. The other genes remained relatively constant during the entire culture duration.

6.4.3 Isolation of Promoters from Dynamic Genes

For each of the three genes identified above (Mmp12, Txnip, and Serpinf1), a fragment of approximate 1000 bp was isolated from Chinese hamster liver genomic DNA. These fragments contain part of the 5' UTR and part of the upstream region of the putative transcription start site (TSS). A representation of the Txnip promoter fragment is shown in Figure 37a. The upstream region of Txnip promoter contains a number of classic elements such as a TATA box and two inverted CAT boxes. In addition, it also contains other regulatory elements including a carbohydrate response element (ChoRE) and binding sites of FOXO and several other transcription factors.

These promoter fragments were cloned upstream of a fusion gene composed of a blasticidin resistance (BSD) marker and an enhanced green fluorescent protein (EGFP) gene as shown in Figure 37b. The resulting plasmids were digested with either ScaI (single digest) or a combination of ScaI, AatII, and XhoI (triple digest) prior to transfection. The triple digest was designed such that the DNA molecule used for transfection contained no elements other than the promoter of interest, the fusion transgene, and the polyadenylation signal. This design prevented any possible cryptic transcriptional regulatory elements from having unwanted effects on the expression of the transgenes. Stable pools and clones were isolated following transfection and selection in blasticidin.

6.4.4 Characterization of Expression Profiles of Transgenes Driven by Txnip Promoter

The presence of the transgenes in the genomes of stable pools and clones was subsequently confirmed by genomic DNA PCR with primers flanking both BSD and EGFP genes. A 421-bp PCR product was found in all selected pools and clones as well as the positive control (pTet_BSD_EGFP plasmid) but not in the negative control (non-transfected cells' genomic DNA). Figure 37c shows an example of stable pools and clones harboring the pTxnip_BSD_EGFP plasmid.

Fed-batch cultures were performed on the isolated stable pool and three randomly chosen clones for each plasmid construct. Samples were collected over different culture times for characterization of the expression trends of the transgenes. As shown in Figure 41a, the pool and all three randomly selected clones of singly digested pTxnip_BSD_EGFP consistently displayed the expected dynamic expression profiles of both BSD and EGFP. The BSD gene was expressed 4- to 16-fold higher in the late-stationary phase compared to the mid-exponential phase. The increase in the EGFP expression level was somewhat less significant; ranging from 2- to 8-fold in the pool and clones. As expected, the endogenous Txnip gene was highly dynamic, with expression increasing between 4- and 16-fold.

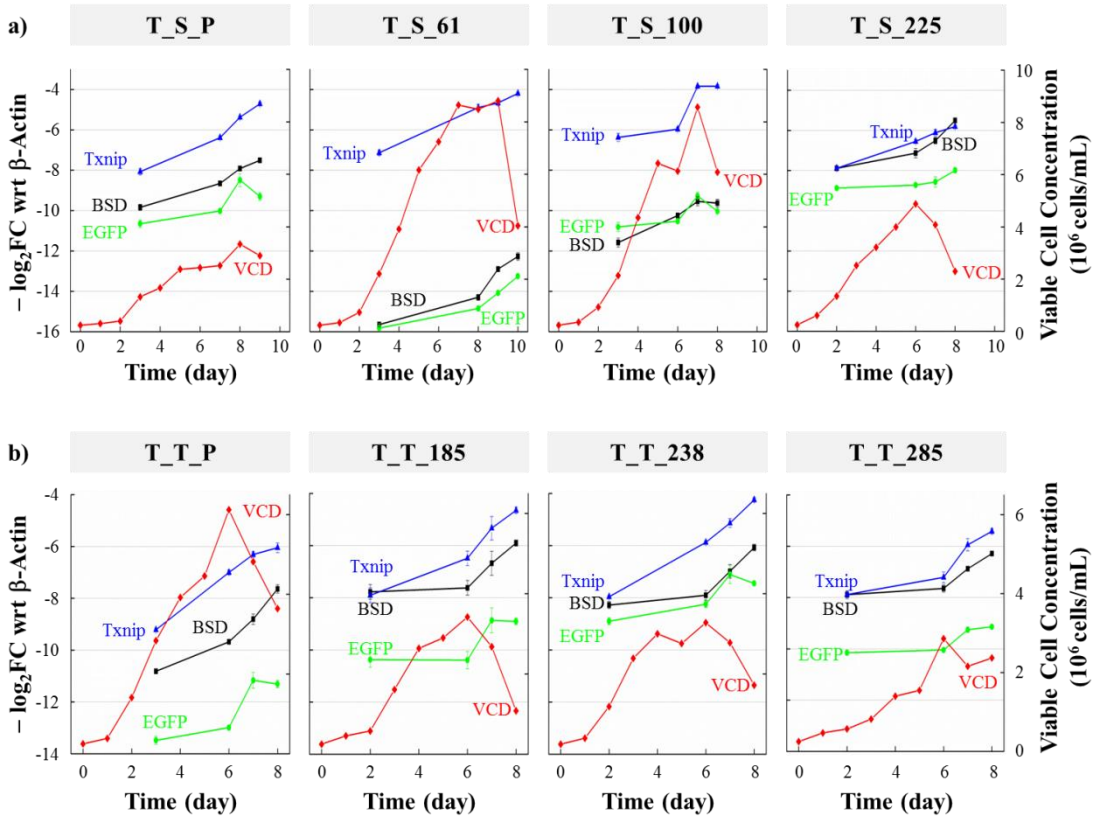


Figure 41: Expression time profiles of the transgenes (BSD in black and EGFP in green) and the endogenous gene (Txnip in blue) in fed-batch cultures of stable pools and clones harboring pTxnip_BSD_EGFP.

a) pTxnip_BSD_EGFP was digested with ScaI prior to transfection
b) pTxnip_BSD_EGFP was digested with ScaI, AatII, and XhoI prior to transfection

The viable cell concentration in each culture was also shown in red for reference.

Although both transgenes demonstrated the expected dynamic trend, they were not expressed at the same level. In most cases, EGFP expression level was lower than that of BSD. In some clones, the difference between BSD and EGFP expression levels was almost negligible (T_S_61, T_S_100, and T_T_238) whereas in other clones, this difference could be over 4-fold.

Furthermore, both transgenes was expressed at lower levels relative to the endogenous gene Txnip. In one clone (T_S_225), the expression of the BSD gene was only slightly lower than that of Txnip. However, in another clone (T_S_61), this difference increased to as much as 64-fold.

In the pool and all three randomly selected clones harboring the triply digested pTxnip_BSD_EGFP, both transgenes and the endogenous gene also exhibited expression dynamics (Figure 41b). As the cells entered the late-stationary phase, they expressed BSD 4- to 16-fold and EGFP 4- to 8-fold higher compared to the mid-exponential phase. Despite having the same expression trend, a difference of up to 8-fold between BSD and EGFP expression levels was observed. Furthermore, BSD was expressed at slightly lower levels than Txnip (less than 3-fold difference).

6.4.5 Characterization of Expression Profiles of Transgenes Driven by Mmp12 Promoter

When the promoter fragment of Mmp12 gene was used to drive the expression of the transgenes, virtually no dynamic trend was observed in stable pools regardless of the restriction digest type (Figure 42). Among the three randomly selected clones for each type of restriction digest, only one appeared to express BSD and EGFP in the expected dynamic manner over the course of the fed-batch cultures. In clone M_S_245 of the single digest, both transgenes were expressed 4-fold higher in the late stationary phase relative to the mid exponential phase. In clone M_T_54 of the triple digest, this increase was up to 16-fold.

Similar to the clones harboring pTxnip_BSD_EGFP plasmid, most clones expressing BSD_EGFP under the control of an Mmp12 promoter also displayed a difference in BSD and EGFP expression levels. In most cases, this difference could be up to 4-fold. However, in clone M_S_245, the expression of EGFP was slightly higher than that of BSD.

Interestingly, the expression levels of the transgenes were not always lower than that of the endogenous gene Mmp12 in all clones. In two clones M_S_235 and M_T_54, BSD was expressed from 4- to 64-fold higher than Mmp12. This trend was reverted in other clones. These opposite trends in the clones could explain the cross-over between BSD and Mmp12 expression levels in the pools, which were supposedly an unbiased mixture of the clones.

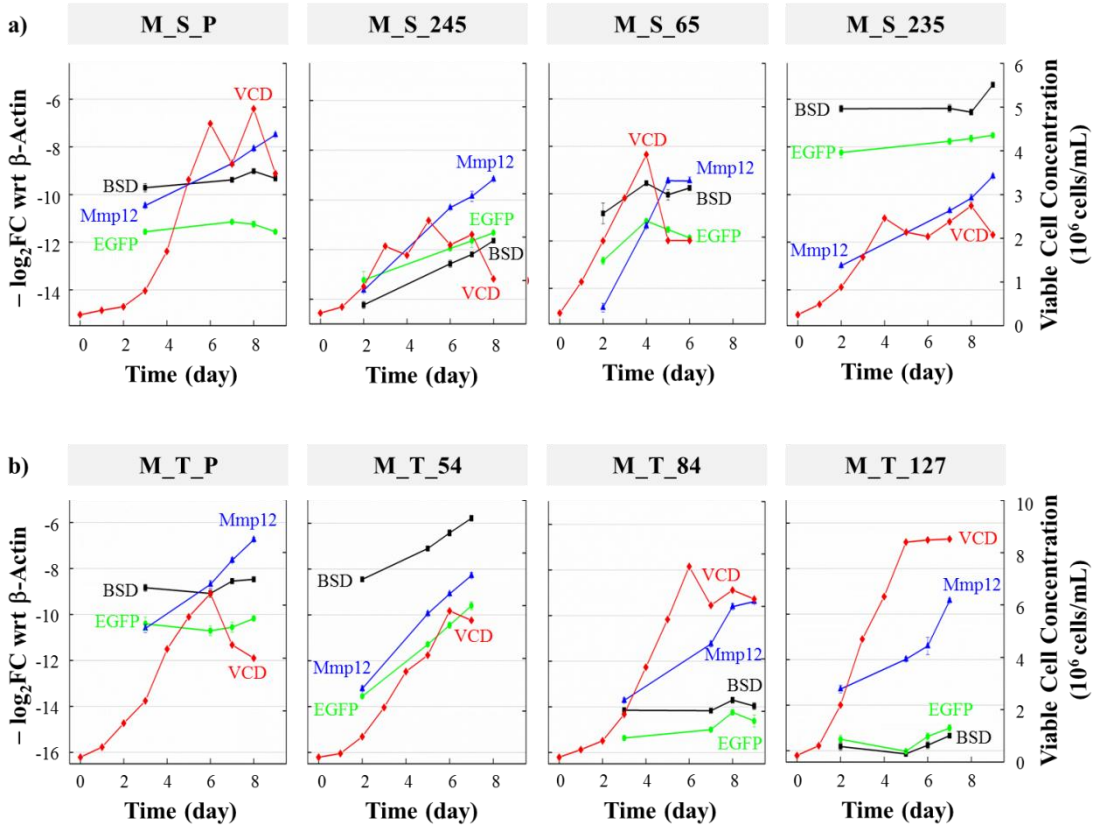


Figure 42: Expression time profiles of the transgenes (BSD in black and EGFP in green) and the endogenous gene (Mmp12 in blue) in fed-batch cultures of stable pools and clones harboring pMmp12_BSD_EGFP.

a) pMmp12_BSD_EGFP was digested with ScaI prior to transfection
b) pMmp12_BSD_EGFP was digested with ScaI, AatII, and XhoI prior to transfection

The viable cell concentration in each culture was also shown in red for reference.

6.4.6 Characterization of Expression Profiles of Transgenes Driven by Serpinf1 Promoter

Among all stable pools and clones harboring the pSerpinf1_BSD_EGFP plasmid, only one clone displayed the expected dynamic trend of the transgenes (S_T_72) (Figure 43). In this clone, both BSD and EGFP genes were expressed nearly16 fold higher in the stationary phase compared to the mid-exponential phase. In other clones and pools, BSD and EGFP expression levels were relatively unchanged, or even decreasing, with time.

Even though both transgenes were expressed with similar trends, there was still a difference between their expression levels. In most cases, EGFP was expressed at much

lower levels with respect to BSD. Typically, a difference of up to 4-fold could be observed.

It was surprising to see that the endogenous gene, Serpinf1, was expressed at various levels in different clones and pools. The $\log_2$ fold-change of Serpinf1 relative to β-Actin ranged from 12 to 8. Furthermore, the difference between Serpinf1 and BSD expression levels also varied significantly across clones. In some cases, Serpinf1 was expressed up to 30-fold lower than was BSD (S_S_194) whereas in others, Serpinf1 was expressed up to 16-fold higher (S_T_P).

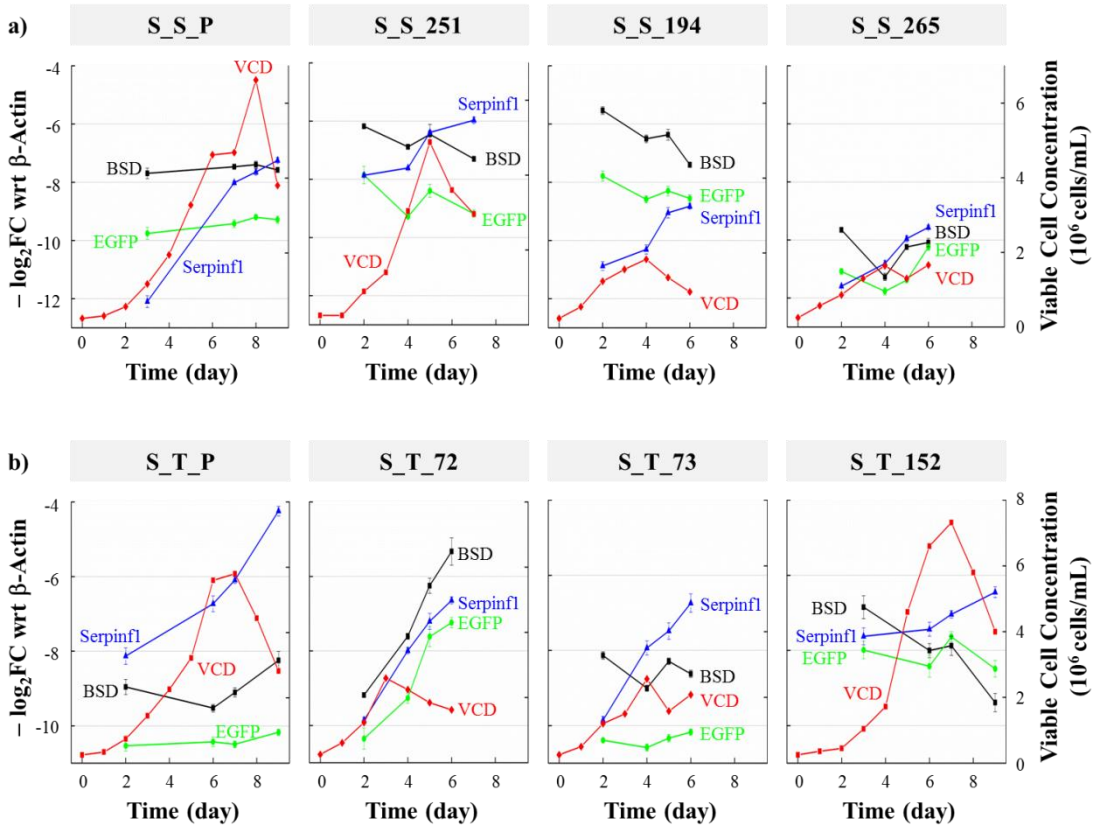


Figure 43: Expression time profiles of the transgenes (BSD in black and EGFP in green) and the endogenous gene (Serpinf1 in blue) in fed-batch cultures of stable pools and clones harboring pSerpinf1_BSD_EGFP.

a) pSerpinf1_BSD_EGFP was digested with ScaI prior to transfection
b) pSerpinf1_BSD_EGFP was digested with ScaI, AatII, and XhoI prior to transfection

The viable cell concentration in each culture was also shown in red for reference.

6.5 DISCUSSION

To our knowledge, this is the first study which presents the novel concept of dynamic cell engineering. A historical data set from 72 time-course microarrays was mined to uncover thousands of genes with various expression trends over the course of a typical fed-batch culture of an industrial CHO cell line. These genes encompass a wide range of normalized intensities, from a few hundred to tens of thousands. The large number of combinations of trends and intensities can arguably create a rich repertoire of expression dynamics. If promoter regions of these genes are identified and isolated, they can potentially be used to drive the expression of transgenes following the desired dynamic trends and levels.

Time-dependent expression dynamics has also been discussed in a few previously published reports. In a study of transcriptional responses in CHO cells under the effect of temperature shift and sodium buryrate treatment, hundreds to thousands of genes were either up- or down-regulated in the late stage of the fed-batch culture relative to the exponential phase (Kantardjieff et al. 2010). A slightly smaller number of genes were identified in the untreated culture at 37^{o}C. Among the genes most strongly induced by butyrate treatment at 33^{o}C, Txnip was expressed at more than 23-fold increase over the course of the culture. In another study, the time dynamics of regulatory motifs that control growth arrest and differentiation of a human monocytic cell line were investigated (Suzuki et al. 2009). At least 30 of such motifs have been identified and clustered into 9 groups with similar time dynamics: three of up-regulated trend, three of down-regulated trend, and three of mixed transient dynamics.

Analysis in a functional context revealed distinct cellular functions which are enriched within each of the six dynamic clusters identified in our study. Genes involved in regulation of apoptosis, redox balance, and protein secretion appear to be expressed at increasing levels over the course of a typical fed-batch culture. In contrast, the expression of genes related to cell cycle, mRNA processing, and ribosome biosynthesis gradually decreases over the same time period. Among genes with increasing expression trends, a small number of genes (15) with sufficient genomic resources were further analyzed. Only three of them, namely Mmp12, Txnip, and Serpinf1, demonstrated

expected dynamic trends in a second CHO cell line, possibly due to variability among different cell lines and culture conditions.

A Txnip promoter fragment of approximately 800 bp was successfully isolated and used to drive the expression of a fusion gene of BSD and EGFP. In stable pools and clones, this promoter was shown to be capable of driving the expression of this fusion gene following the expected trend. However, the fusion gene was expressed at lower levels compared to the endogenous gene of Txnip in most cases. This result indicates that the isolated promoter fragment, albeit capable of conferring essential transcriptional activities, may not harbor all necessary regulatory elements to fully mimic the endogenous expression. Furthermore, the expression levels of the fusion gene varied significantly across different clones. This is likely due to the influence of chromosomal integration site as elaborated in literature (Dorer and Henikoff 1994; Festenstein et al. 1996; Kleinjan and van Heyningen 1998).

The next step following this proof-of-concept study would be using this promoter to control the dynamics of target genes with significant cellular functions. Potential targets for dynamic control include a number of cellular processes which are known to be critical for the production of secretory protein products, such as apoptosis and energy metabolism. Apoptosis is one of the major contributors to cell death upon exposure to the high-stress environment in bioreactors experienced by most industrial cell lines (Arden 2004). Inhibition of apoptosis has been shown to prolong culture duration, resulting in higher product titer (Figueroa et al. 2007). In addition, energy metabolism is known to impose profound effects on cell growth and productivity as reviewed recently (Mulukutla et al. 2010). Modulating energy metabolism to reduce lactate production has been shown to significantly increase product titer (Zhou et al. 2011). Ideally, implementation of such strategies would be desirable in the late stages of the culture, when environmental stresses including lactate concentration approach critical levels. This aim could be achieved using a dynamic endogenous promoter with low activity in the mid-exponential phase and gradually increasing activity towards the late-stationary phase, such as that of Txnip. Thus the results in this study have demonstrated the feasibility of modulating cellular behaviors in a dynamic manner.

The increasing availability of mammalian genome sequences in recent years is opening up a remarkable opportunity to explore and gain better understanding of transcription regulation. In the past few years, various ENCODE and FANTOM projects have shed light on previously unknown complexities of mammalian transcription networks (The ENCODE Consortium 2007; The FANTOM Consortium 2005). More regulatory mechanisms with a multitude of cis- and trans-acting elements continue to be uncovered. Although substantial efforts need to be invested to fully understand the depth and breadth of gene regulation, harnessing this knowledge will ultimately lead to the ability to modulate gene expression in unimaginable ways.

7 CONCLUSION AND FUTURE DIRECTIONS

High-throughput genomic and process analytical technologies hold the potential to bring fundamental understanding of cellular changes occurring at each and every step along a cell culture process. This thesis describes the application of advanced multivariate approaches in uncovering valuable information from the immense volume of data generated by these technologies. The insights gained through this analysis continue to broaden our understanding of multiple facets of mammalian cell culture. Furthermore, they offer great opportunities for enhancement of process performance.

Ceaseless efforts have been made by the entire scientific community to develop genomic resources for a large number of mammalian species as well as required analysis approaches. The results yielded by these fundamental studies will accelerate the product development pipeline in the pharmaceutical industry. Furthermore, as more data of multiple types are accumulating, we could anticipate a swift move from the current exploratory phase to more rigorous evidence-based intervention. Integrating these conclusions will certainly increase our capability to discover novel means to transform the current model of pharmaceutical research and development. Such transformational changes are certainly needed to bring back the innovation level that prevailed in the era of blockbusters.

8 REFERENCES

Aggarwal S. 2011. What's fueling the biotech engine—2010 to 2011. Nature Biotechnology 29(12):1083-1089.

Aiba K, Sharov AA, Carter MG, Foroni C, Vescovi AL, Ko MSH. 2006. Defining a Developmental Path to Neural Fate by Global Expression Profiling of Mouse Embryonic Stem Cells and Adult Neural Stem/Progenitor Cells. STEM CELLS 24(4):889-895.

Altamirano C, Illanes A, Becerra S, Cairó JJ, Gòdia F. 2006. Considerations on the lactate consumption by CHO cells in the presence of galactose. Journal of Biotechnology 125(4):547-556.

Alter O, Brown PO, Botstein D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences 97(18):10101-10106.

Ambrosi DJ, Tanasijevic B, Kaur A, Obergfell C, O'Neill RJ, Krueger W, Rasmussen TP. 2007. Genome-Wide Reprogramming in Hybrids of Somatic Cells and Embryonic Stem Cells. STEM CELLS 25(5):1104-1113.

Angelini C, Cutillo L, De Canditiis D, Mutarelli M, Pensky M. 2008. BATS: a Bayesian user-friendly software for analyzing time series microarray experiments. BMC Bioinformatics 9:415.

Angelini C, De Canditiis D, Mutarelli M, Pensky M. 2007. A Bayesian approach to estimation and testing in time-course microarray experiments. Stat Appl Genet Mol Biol 6:Article24.

Anichini A, Scarito A, Molla A, Parmiani G, Mortarini R. 2003. Differentiation of CD8+ T Cells from Tumor-Invaded and Tumor-Free Lymph Nodes of Melanoma Patients: Role of Common Î³-Chain Cytokines. The Journal of Immunology 171(4):2134-2141.

Arden N. 2004. Life and death in mammalian cell culture: strategies for apoptosis inhibition. Trends in Biotechnology 22(4):174-180.

Aryee DNT, Niedan S, Kauer M, Schwentner R, Bennani-Baiti IM, Ban J, Muehlbacher K, Kreppel M, Walker RL, Meltzer P and others. 2010. Hypoxia Modulates EWS-FLI1 Transcriptional Signature and Enhances the Malignant Properties of Ewing's Sarcoma Cells In vitro. Cancer Research 70(10):4015-4023.

Ayache S, Panelli M, Byrne K, Slezak S, Leitman S, Marincola F, Stroncek D. 2006. Comparison of proteomic profiles of serum, plasma, and modified media supplements used for cell culture and expansion. Journal of Translational Medicine 4(1):40.

Bachinger T, Riese U, Eriksson R, Mandenius C-F. 2000. Monitoring cellular state transitions in a production-scale CHO-cell process using an electronic nose. Journal of Biotechnology 76(1):61-71.

Baker DA, Russell S. 2009. Gene expression during Drosophila melanogaster egg development before and after reproductive diapause. BMC Genomics 10:242.

Baker TK, Carfagna MA, Gao H, Dow ER, Li Q, Searfoss GH, Ryan TP. 2001. Temporal Gene Expression Analysis of Monolayer Cultured Rat Hepatocytes. Chemical Research in Toxicology 14(9):1218-1231.

Bar-Joseph Z. 2004. Analyzing time series gene expression data. Bioinformatics 20(16):2493-503.

Bar-Joseph Z, Gerber GK, Gifford DK, Jaakkola TS, Simon I. 2003. Continuous Representations of Time-Series Gene Expression Data. Journal of Computational Biology 10(3-4):341-356.

Barkow S, Bleuler S, Prelic A, Zimmermann P, Zitzler E. 2006. BicAT: a biclustering analysis toolbox. Bioinformatics 22(10):1282-3.

Barnes LM, Bentley CM, Dickson AJ. 2004. Molecular definition of predictive indicators of stable protein expression in recombinant NS0 myeloma cells. Biotechnology and Bioengineering 85(2):115-121.

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. 2005. Reverse engineering of regulatory networks in human B cells. Nat Genet 37(4):382-90.

Basso K, Saito M, Sumazin P, Margolin AA, Wang K, Lim WK, Kitagawa Y, Schneider C, Alvarez MJ, Califano A and others. 2010. Integrated biochemical and computational approach identifies BCL6 direct target genes controlling multiple pathways in normal germinal center B cells. Blood 115(5):975-84.

Bell SL, Bebbington C, Scott MF, Wardell JN, Spier RE, Bushell ME, Sanders PG. 1995. Genetic engineering of hybridoma glutamine metabolism. Enzyme and Microbial Technology 17(2):98-106.

Bendig M. 1988. The production of foreign proteins in mammalian cells. Genetic Engineering 7:91-127.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57(1):289-300.

Bezdek J. 1981. Pattern Recognition with Fuzzy Objective Function Algorithms (Advanced Applications in Pattern Recognition): Springer.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19(2):185-193.

Bonfanti P, Claudinot S, Amici AW, Farley A, Blackburn CC, Barrandon Y. 2010. Microenvironmental reprogramming of thymic epithelial cells to skin multipotent stem cells. Nature 466(7309):978-982.

Bonferroni CE. 1936. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8:3-62.

Boser BE, Guyon IM, Vapnik VN. 1992. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory. Pittsburgh, Pennsylvania, United States: ACM.

Boutros PC, Okey AB. 2005. Unsupervised pattern recognition: An introduction to the whys and wherefores of clustering microarray data. Briefings in Bioinformatics 6(4):331-343.

Breiman L. 2001. Random Forests. Machine Learning 45(1):5-32.

Breiman L, Friedman J, Olshen R, Stone C. 1984. Classification and Regression Trees: Wadsworth International Group.

Broad Institute. 2011. Mesocricetus auratus Genome sequencing.

Brunet J-P, Tamayo P, Golub TR, Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences of the United States of America 101(12):4164-4169.

Brusniak MY, Bodenmiller B, Campbell D, Cooke K, Eddes J, Garbutt A, Lau H, Letarte S, Mueller LN, Sharma V and others. 2008. Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. BMC Bioinformatics 9:542.

Buck KKS, Subramanian V, Block DE. 2002. Identification of Critical Batch Operating Parameters in Fed-Batch Recombinant E. coli Fermentations Using Decision Tree Analysis. Biotechnology Progress 18(6):1366-1376.

Butler JEF. 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes & Development 16(20):2583-2592.

Butte AJ, Kohane IS. 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput:418-29.

Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. 2000. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci U S A 97(22):12182-6.

Carey VJ, Gentry J, Whalen E, Gentleman R. 2005. Network structures and algorithms in Bioconductor. Bioinformatics 21(1):135-6.

Castro-Melchor M, Charaniya S, Karypis G, Takano E, Hu W-S. 2010. Genome-wide inference of regulatory networks in Streptomyces coelicolor. BMC Genomics 11(1):578.

Chang C-C, Lin C-J. 2001. LIBSVM : a library for support vector machines.

Charaniya S, Hu W, Karypis G. 2008. Mining bioprocess data: opportunities and challenges. Trends in Biotechnology 26(12):690-699.

Charaniya S, Karypis G, Hu W-S. 2009. Mining transcriptome data for function–trait relationship of hyper productivity of recombinant antibody. Biotechnology and Bioengineering 102(6):1654-1669.

Charaniya S, Le H, Rangwala H, Mills K, Johnson K, Karypis G, Hu W-S. 2010. Mining manufacturing data for discovery of high productivity process characteristics. Journal of Biotechnology 147(3–4):186-197.

Chen K, Liu Q, Xie L, Sharp PA, Wang DIC. 2001. Engineering of a mammalian cell line for reduction of lactate formation and high monoclonal antibody production. Biotechnology and Bioengineering 72(1):55-61.

Cheng KO, Law NF, Siu WC, Lau TH. 2007. BiVisu: software tool for bicluster detection and visualization. Bioinformatics 23(17):2342-4.

Cheng Y, Church GM. 2000. Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol 8:93-103.

Choi YL, Tsukasaki K, O'Neill MC, Yamada Y, Onimaru Y, Matsumoto K, Ohashi J, Yamashita Y, Tsutsumi S, Kaneda R and others. 2006. A genomic analysis of adult T-cell leukemia. Oncogene 26(8):1245-1255.

Chong I-G, Jun C-H. 2005. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems 78(1-2):103-112.

Chong WPK, Goh LT, Reddy SG, Yusufi FNK, Lee DY, Wong NSC, Heng CK, Yap MGS, Ho YS. 2009. Metabolomics profiling of extracellular metabolites in recombinant Chinese Hamster Ovary fed-batch culture. Rapid Communications in Mass Spectrometry 23(23):3763-3771.

Chusainow J, Yang YS, Yeo JHM, Toh PC, Asvadi P, Wong NSC, Yap MGS. 2009. A study of monoclonal antibody-producing CHO cell lines: What makes a stable high producer? Biotechnology and Bioengineering 102(4):1182-1196.

Ciaccio MF, Wagner JP, Chuu CP, Lauffenburger DA, Jones RB. 2010. Systems analysis of EGF receptor signaling dynamics with microwestern arrays. Nat Methods 7(2):148-55.

Coleman MC, Block DE. 2006. Retrospective optimization of time-dependent fermentation control strategies using time-independent historical data. Biotechnology and Bioengineering 95(3):412-423.

Conesa A, Nueda MJ, Ferrer A, Talon M. 2006. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. Bioinformatics 22(9):1096-102.

Cover T, Hart P. 1967. Nearest neighbor pattern classification. Information Theory, IEEE Transactions on 13(1):21-27.

Cruz HJ, Moreira JL, Carrondo MJT. 1999. Metabolic shifts by nutrient manipulation in continuous cultures of BHK cells. Biotechnology and Bioengineering 66(2):104-113.

Czernicki T, Zegarska J, Paczek L, Cukrowska B, Grajokowska W, Zajaczkowska A, Brudzewski K, Ulaczyk J, Marchel A. 2007. Gene expression profile as a prognostic factor in high-grade gliomas. International Journal of Oncology 30:55-64.

Dahlquist KD. 2002. Using GenMAPP and MAPPFinder to View Microarray Data on Biological Pathways and Identify Global Trends in the Data: John Wiley & Sons, Inc.

Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 31(1):19-20.

De Bruyne V, Al-Mulla F, Pot B. 2005. Methods for Microarray Data Analysis: Humana Press.

De Hoon MJ, Imoto S, Miyano S. 2002. Statistical analysis of a small set of time-ordered gene expression data using linear splines. Bioinformatics 18(11):1477-85.

de Jong S. 1993. SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 18(3):251-263.

De Leon Gatti M, Wlaschin KF, Nissom PM, Yap M, Hu W-S. 2007. Comparative transcriptional analysis of mouse hybridoma and recombinant Chinese hamster ovary cells undergoing butyrate treatment. Journal of Bioscience and Bioengineering 103(1):82-91.

Dembele D, Kastner P. 2003. Fuzzy C-means method for clustering microarray data. Bioinformatics 19(8):973-980.

Do JH, Choi D-K. 2008. Clustering approaches to identifying gene expression patterns from DNA microarray data. Molecules and cells 25(2):279-288.

Dojer N, Gambin A, Mizera A, Wilczynski B, Tiuryn J. 2006. Applying dynamic Bayesian networks to perturbed gene expression data. BMC Bioinformatics 7(1):249.

Doniger S, Salomonis N, Dahlquist K, Vranizan K, Lawlor S, Conklin B. 2003. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biology 4(1):R7.

Dopazo J, Zanders E, Dragoni I, Amphlett G, Falciani F. 2001. Methods and approaches in the analysis of gene expression data. Journal of Immunological Methods 250(1-2):93-112.

Dorai H, Kyung YS, Ellis D, Kinney C, Lin C, Jan D, Moore G, Betenbaugh MJ. 2009. Expression of anti-apoptosis genes alters lactate metabolism of Chinese Hamster Ovary cells in culture. Biotechnology and Bioengineering 103(3):592-608.

Dorer DR, Henikoff S. 1994. Expansions of transgene repeats cause heterochromatin formation and gene silencing in Drosophila. Cell 77(7):993-1002.

Efron B, Tibshirani R. 2007. On testing the significance of sets of genes. The Annals of Applied Statistics 1(1):107-129.

Ernst J, Bar-Joseph Z. 2006. STEM: a tool for the analysis of short time series gene expression data. BMC Bioinformatics 7(1):191.

Everitt B. 1974a. Cluster Analysis. London: Heinemann Education Books.

Everitt BS. 1974b. Cluster analysis. London: Heinemann Educational [for] the Social Science Research Council.

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5(1):e8.

Fann CH, Guirgis F, Chen G, Lao MS, Piret JM. 2000. Limitations to the amplification and stability of human tissue-type plasminogen activator expression by Chinese hamster ovary cells. Biotechnol Bioeng 69(Copyright (C) 2012 U.S. National Library of Medicine.):204-12.

Feldser DM, Kostova KK, Winslow MM, Taylor SE, Cashman C, Whittaker CA, Sanchez-Rivera FJ, Resnick R, Bronson R, Hemann MT and others. 2010. Stage-specific sensitivity to p53 restoration during lung cancer progression. Nature 468(7323):572-575.

Festenstein R, Tolaini M, Corbella P, Mamalaki C, Parrington J, Fox M, Miliou A, Jones M, Kioussis D. 1996. Locus control region function and heterochromatin-induced position effect variegation. Science (New York, N.Y.) 271(5252):1123-5.

Figueroa B, Ailor E, Osborne D, Hardwick JM, Reff M, Betenbaugh MJ. 2007. Enhanced cell culture performance using inducible anti-apoptotic genes E1B-19K and Aven in the production of a monoclonal antibody with Chinese hamster ovary cells. Biotechnology and Bioengineering 97(4):877-892.

Figueroa B, Chen S, Oyler GA, Hardwick JM, Betenbaugh MJ. 2004. Aven and Bcl-xL enhance protection against apoptosis for mammalian cells exposed to various culture conditions. Biotechnology and Bioengineering 85(6):589-600.

Finn GK, Kurz BW, Cheng RZ, Shmookler Reis RJ. 1989. Homologous plasmid recombination is elevated in immortally transformed cells. Molecular and Cellular Biology 9(9):4009-4017.

Fogolín MB, Wagner R, Etcheverrigaray M, Kratje R. 2004. Impact of temperature reduction and expression of yeast pyruvate carboxylase on hGM-CSF-producing CHO cells. Journal of Biotechnology 109(1–2):179-191.

Fortier JM, Payton JE, Cahan P, Ley TJ, Walter MJ, Graubert TA. 2010. POU4F1 is associated with t(8;21) acute myeloid leukemia and contributes directly to its unique transcriptional signature. Leukemia 24(5):950-957.

Frades I, Matthiesen R. 2010. Overview on Techniques in Cluster Analysis. Bioinformatics methods in clinical research 593:81-107.

Frigyesi A, Hoglund M. 2008. Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. Cancer Informatics 6:275-292.

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer Ml, Haussler D. 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10):906-914.

Gaffney S, Smyth P. 2005. Joint Probabilistic Curve Clustering and Alignment. Advances in Neural Information Processing Systems.

Gandor C, Leist C, Fiechter A, Asselbergs FAM. 1995. Amplification and expression of recombinant genes in serum-independent Chinese hamster ovary cells. FEBS Letters 377(3):290-294.

Gene Set Enrichment Analysis. http://www.broadinstitute.org/gsea/index.jsp.

Glacken MW, Fleischaker RJ, Sinskey AJ. 1986. Reduction of waste product excretion via nutrient control: Possible strategies for maximizing product and cell yields on serum in cultures of mammalian cells. Biotechnology and Bioengineering 28(9):1376-1389.

Gollub J, Sherlock G, Alan K, Brian O. 2006. [10] Clustering Microarray Data. Methods in Enzymology: Academic Press. p 194-213.

Goncalves JP, Madeira SC, Oliveira AL. 2009. BiGGEsTS: integrated environment for biclustering analysis of time series gene expression data. BMC Res Notes 2:124.

Gunther JC, Conner JS, Seborg DE. 2007. Fault Detection and Diagnosis in an Industrial Fed-Batch Cell Culture Process. Biotechnology Progress 23(4):851-857.

Hammill L, Welles J, Carson GR. 2000. The gel microdrop secretion assay: Identification of a low productivity subpopulation arising during the production of human antibody in CHO cells. Cytotechnology 34(1):27-37.

Han X. 2008. Improving gene expression cancer molecular pattern discovery using nonnegative principal component analysis. Genome Informatics 21:200-211.

Harris M. 1984. Variants inducible for glutamine synthetase in V79-56 cells. Somatic Cell and Molecular Genetics 10(3):275-281.

Heard NA, Holmes CC, Stephens DA. 2006. A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes. Journal of the American Statistical Association 101(473):18-29.

Heckerman D. 1998. A tutorial on learning with Bayesian networks. In: Jordan MI, editor. Learning in Graphical Models. Boston: Kluwer Academic.

Hoogerwerf WA, Sinha M, Conesa A, Luxon BA, Shahinian VB, Cornelissen G, Halberg F, Bostwick J, Timm J, Cassone VM. 2008. Transcriptional profiling of mRNA expression in the mouse distal colon. Gastroenterology 135(6):2019-29.

Hu W, Dodge T, Frame K, Himes V. 1987. Effect of glucose on the cultivation of mammalian cells. Developments in Biological Standardization 66:279-90.

Ihnen M, Wirtz RM, Kalogeras KT, Milde-Langosch K, Schmidt M, Witzel I, Eleftheraki AG, Papadimitriou C, Janicke F, Briassoulis E and others. 2010. Combination of osteopontin and activated leukocyte cell adhesion molecule as potent prognostic discriminators in HER2- and ER-negative breast cancer. Br J Cancer 103(7):1048-1056.

Irani N, Wirth M, van den Heuvel J, Wagner R. 1999. Improvement of the primary metabolism of cell cultures by introducing a new cytoplasmic pyruvate carboxylase reaction. Biotechnology and Bioengineering 66(4):238-246.

Jacob NM, Yusufi FNK, Chin JX, Lee TS, Le H, Johnson KC, Vishwanathan N, Loo B, Ramaraj T, Retzel EF and others. Sequencing the Chinese hamster genome as a resource for CHO genome engineering (in preparation).

Jacob NM, Yusufi FNK, Chin JX, Lee TS, Le H, Johnson KC, Vishwanathan N, Loo B, Ramaraj T, Retzel EF and others. Sequencing the Chinese hamster genome as a resource for CHO genome engineering (in preparation).

Jansen JJ, Hoefsloot HCJ, Greef Jvd, Timmerman ME, Westerhuis JA, Smilde AK. 2005. ASCA: analysis of multivariate data obtained from an experimental design. Journal of Chemometrics 19(9):469-481.

Jayapal K, Wlaschin K, Hu W-S, Yap M. 2007. Recombinant Protein Therapeutics from Cho Cells - 20 Years and Counting. CHO Consortium: SBE Special Edition:40-47.

Jiang W, Li X, Rao S, Wang L, Du L, Li C, Wu C, Wang H, Wang Y, Yang B. 2008. Constructing disease-specific gene networks using pair-wise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements. BMC Syst Biol 2:72.

Jolliffe I. 2005. Principal Component Analysis: John Wiley & Sons, Ltd.

Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. Nature 468(7325):811-814.

Kallel H, Jouini A, Majoul S, Rourou S. 2002. Evaluation of various serum and animal protein free media for the production of a veterinary rabies vaccine in BHK-21 cells. Journal of Biotechnology 95(3):195-204.

Kantardjieff A, Jacob NM, Yee JC, Epstein E, Kok Y-J, Philp R, Betenbaugh M, Hu W-S. 2010. Transcriptome and proteome analysis of Chinese hamster ovary cells under low temperature and butyrate treatment. Journal of Biotechnology 145(2):143-159.

Kassidas A, MacGregor JF, Taylor PA. 1998. Synchronization of batch trajectories using dynamic time warping. AIChE Journal 44(4):864-875.

Kerr MK, Churchill GA. 2001. Statistical design and the analysis of gene expression microarray data. Genet Res 77(2):123 - 8.

Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C and others. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7(6):673-679.

Kim E, Kim N, Lee G. 1998. Development of a serum-free medium for the production of humanized antibody from chinese hamster ovary cells using a statistical design. In Vitro Cellular & Developmental Biology - Animal 34(10):757-761.
Kim S, Lee G. 2007a. Down-regulation of lactate dehydrogenase-A by siRNAs for reduced lactic acid formation of Chinese hamster ovary cells producing thrombopoietin. Applied Microbiology and Biotechnology 74(1):152-159.
Kim S, Lee G. 2007b. Functional expression of human pyruvate carboxylase for reduced lactic acid formation of Chinese hamster ovary cells (DG44). Applied Microbiology and Biotechnology 76(3):659-665.
Kim S, Lee J, Bae J. 2006. Effect of data normalization on fuzzy clustering of DNA microarray data. BMC Bioinformatics 7(1):134.
Kim SH, Lee GM. 2006. Down-regulation of lactate dehydrogenase-A by siRNAs for reduced lactic acid formation of Chinese hamster ovary cells producing thrombopoietin. Applied Microbiology and Biotechnology 74(1):152-159.
Kim SH, Lee GM. 2007c. Functional expression of human pyruvate carboxylase for reduced lactic acid formation of Chinese hamster ovary cells (DG44). Applied Microbiology and Biotechnology 76(3):659-665.
Kim SY, Imoto S, Miyano S. 2003. Inferring gene networks from time series microarray data using dynamic Bayesian networks. Brief Bioinform 4(3):228-235.
Kingsford C, Salzberg SL. 2008. What are decision trees? Nat Biotech 26(9):1011-1013.
Kirdar AO, Green KD, Rathore AS. 2008. Application of Multivariate Data Analysis for Identification and Successful Resolution of a Root Cause for a Bioprocessing Application. Biotechnology Progress 24(3):720-726.
Kleinjan D-J, van Heyningen V. 1998. Position Effect in Human Genetic Disease. Human Molecular Genetics 7(10):1611-1618.
Kohonen T. 2001. Self-Organizing Maps. Berlin: Springer.
Korke R, Gatti MdL, Lau ALY, Lim JWE, Seow TK, Chung MCM, Hu W-S. 2004. Large scale gene expression profiling of metabolic shift of mammalian cells in culture. Journal of Biotechnology 107(1):1-17.
Krampe B, Swiderek H, Al-Rubeai M. 2008. Transcriptome and proteome analysis of antibody-producing mouse myeloma NS0 cells cultivated at different cell densities in perfusion culture. Biotechnology and Applied Biochemistry 50(3):133.
Krogh A. 2008. What are artificial neural networks? Nat Biotech 26(2):195-197.
Laiho P, Kokko A, Vanharanta S, Salovaara R, Sammalkorpi H, Jarvinen H, Mecklin JP, Karttunen TJ, Tuppurainen K, Davalos V and others. 2006. Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. Oncogene 26(2):312-320.
Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW. 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proceedings of the National Academy of Sciences of the United States of America 94(24):13057-13062.
Le A, Cooper CR, Gouw AM, Dinavahi R, Maitra A, Deck LM, Royer RE, Vander Jagt DL, Semenza GL, Dang CV. 2010. Inhibition of lactate dehydrogenase A induces oxidative stress and inhibits tumor progression. Proceedings of the National Academy of Sciences 107(5):2037-2042.
Le H, Castro-Melchor M, Hakemeyer C, Jung C, Szperalski B, Karypis G, Hu W-S. Mining bioprocess data for discovery of key parameters influencing high productivity and quality. Proceedings of the 22nd Annual Meeting of the European Society for Animal Cell Technology (ESACT), Vienna, Austria, May 15-18, 2011. Springer.
Leader B, Baca QJ, Golan DE. 2008. Protein therapeutics: a summary and pharmacological classification. Nat Rev Drug Discov 7(1):21-39.

Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788-791.
Lee YY, Wong KTK, Nissom PM, Wong DCF, Yap MGS. 2007. Transcriptional profiling of batch and fed-batch protein-free 293-HEK cultures. Metabolic Engineering 9(1):52-67.
Leek JT, Monsen E, Dabney AR, Storey JD. 2006. EDGE: extraction and analysis of differential gene expression. Bioinformatics 22(4):507 - 8.
Leung E, Bushel PR. 2006. PAGE: phase-shifted analysis of gene expression. Bioinformatics 22(3):367-8.
Levin AM, de Vries RP, Conesa A, de Bekker C, Talon M, Menke HH, van Peij NN, Wosten HA. 2007. Spatial differentiation in the vegetative mycelium of Aspergillus niger. Eukaryot Cell 6(12):2311-22.
Li P, Zhang C, Perkins E, Gong P, Deng Y. 2007a. Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. BMC Bioinformatics 8(Suppl 7):S13.
Li W, You P, Wei Q, Li Y, Fu X, Ding X, Wang X, Hu Y. 2007b. Hepatic differentiation and transcriptional profile of the mouse liver epithelial progenitor cells (LEPCs) under the induction of sodium butyrate. Front Biosci. 12:1691-8.
Li Y, Wang N, Perkins EJ, Zhang C, Gong P. 2010. Identification and Optimization of Classifier Genes from Multi-Class Earthworm Microarray Dataset. PLoS One 5(10):e13715.
Lian W, Jayapal K, Charaniya S, Mehra S, Glod F, Kyung Y-S, Sherman D, Hu W-S. 2008. Genome-wide transcriptome analysis reveals that a pleiotropic antibiotic regulator, AfsS, modulates nutritional stress response in Streptomyces coelicolor A3(2). BMC Genomics 9(1):56.
Lievens S, Lemmens I, Tavernier J. 2009. Mammalian two-hybrids come of age. Trends in Biochemical Sciences 34(11):579-588.
Liu H, Hiroshi M. 1998. Feature Selection for Knowledge Discovery and Data Mining: Springer. 244 p.
Liu W, Tanasa B, Tyurina OV, Zhou TY, Gassmann R, Liu WT, Ohgi KA, Benner C, Garcia-Bassets I, Aggarwal AK and others. 2010a. PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. Nature 466(7305):508-512.
Liu W, Yuan K, Ye D. 2008. Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis. Journal of Biomedical Informatics 41(4):602-606.
Liu Y, Yang Y, Xu H, Dong X. 2010b. Implication of USP22 in the Regulation of BMI-1, c-Myc, p16INK4a, p14ARF, and Cyclin D2 Expression in Primary Colorectal Carcinomas. Diagnostic Molecular Pathology 19(4):194-200 10.1097/PDM.0b013e3181e202f2.
Lonnstedt I, Britton T. 2005. Hierarchical Bayes models for cDNA microarray gene expression. Biostatistics 6(2):279-291.
Luan Y, Li H. 2003. Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics 19(4):474-82.
Madeira SC, Oliveira AL. 2004. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform 1(1):24-45.
Malphettes L, Freyvert Y, Chang J, Liu P-Q, Chan E, Miller JC, Zhou Z, Nguyen T, Tsai C, Snowden AW and others. 2010. Highly efficient deletion of FUT8 in CHO cell lines using zinc-finger nucleases yields cells that produce completely nonfucosylated antibodies. Biotechnology and Bioengineering 106(5):774-783.
Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. 2006a. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1:S7.
Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A. 2006b. Reverse engineering cellular networks. Nat Protoc 1(2):662-71.

Mehra S, Lian W, Jayapal K, Charaniya S, Sherman D, Hu W-S. 2006. A framework to analyze multiple time series data: A case study with Streptomyces coelicolor. Journal of Industrial Microbiology and Biotechnology 33(2):159-172.

Meyer PE, Kontos K, Lafitte F, Bontempi G. 2007. Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol:79879.

Meyer PE, Lafitte F, Bontempi G. 2008. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics 9:461.

Minsky ML, Papert SA. 1969. Perceptrons: MIT Press.

Moller-Levet CS, Cho KH, Wolkenhauer O. 2003. Microarray data clustering based on temporal variation: FCV with TSD preclustering. Appl Bioinformatics 2(1):35-45.

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E and others. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34(3):267 - 73.

Moriyama M, Hoshida Y, Otsuka M, Nishimura S, Kato N, Goto T, Taniguchi H, Shiratori Y, Seki N, Omata M. 2003. Relevance network between chemosensitivity and transcriptome in human hepatoma cells. Mol Cancer Ther 2(2):199-205.

Morrison DA, Ellis JT. 2003. The Design and Analysis of Microarray Experiments: Applications in Parasitology. DNA and Cell Biology 22(6):357-394.

Mortensen RM, Kingston RE. 2001. Selection of Transfected Mammalian Cells. Current Protocols in Molecular Biology: John Wiley & Sons, Inc.

Mulukutla B, Gramer M, Hu WS. 2011. On metabolic shift to lactate consumption in fed-batch culture of mammalian cells. Metabolic Engineering (accepted).

Mulukutla BC, Khan S, Lange A, Hu W-S. 2010. Glucose metabolism in mammalian cell culture: new insights for tweaking vintage pathways. Trends in Biotechnology 28(9):476-484.

Murphy K, Mian S. 1999. Modelling gene expression data using dynamic Bayesian networks.

Mutskov V, Felsenfeld G. 2004. Silencing of transgene transcription precedes methylation of promoter DNA and histone H3 lysine 9. EMBO J 23(1):138-149.

Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR. 2006. Inference in Bayesian networks. Nat Biotech 24(1):51-53.

Nemenman I, Escola GS, Hlavacek WS, Unkefer PJ, Unkefer CJ, Wall ME. 2007. Reconstruction of metabolic networks from high-throughput metabolite profiling data: in silico analysis of red blood cell metabolism. Ann N Y Acad Sci 1115:102-15.

Noble WS. 2006. What is a support vector machine? Nat Biotech 24(12):1565-1567.

Nueda MJ, Conesa A, Westerhuis JA, Hoefsloot HCJ, Smilde AK, Talon M, Ferrer A. 2007. Discovering gene expression patterns in time course microarray experiments by ANOVA SCA. Bioinformatics 23(14):1792-1800.

Nueda MJ, Sebastian P, Tarazona S, Garcia-Garcia F, Dopazo J, Ferrer A, Conesa A. 2009. Functional assessment of time course microarray data. BMC Bioinformatics 10 Suppl 6:S9.

Nugent R, Meila M. 2010. An overview of clustering applied to molecular biology: Humana Press. 369-404 p.

Oleksiak MF, Churchill GA, Crawford DL. 2002. Variation in gene expression within and among natural populations. Nat Genet 32(2):261-266.

Ozbudak E, Tassy O, Pourquie O. 2010. Spatiotemporal compartmentalization of key physiological processes during muscle precursor differentiation. Proceedings of the National Academy of Sciences 107(9):4224-4229.

Pallavicini MG, DeTeresa PS, Rosette C, Gray JW, Wurm FM. 1990. Effects of methotrexate on transfected DNA stability in mammalian cells. Molecular and Cellular Biology 10(1):401-404.

Pandey G, Yoshikawa K, Hirasawa T, Nagahisa K, Katakura Y, Furusawa C, Shimizu H, Shioya S. 2007. Extracting the hidden features in saline osmotic tolerance in Saccharomyces cerevisiae from DNA microarray data using the self-organizing map: biosynthesis of amino acids. Applied Microbiology and Biotechnology 75(2):415-426.

Parry RM, Jones W, Stokes TH, Phan JH, Moffitt RA, Fang H, Shi L, Oberthuer A, Fischer M, Tong W and others. 2010. k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. Pharmacogenomics J 10(4):292-309.

Pascual L, Blanca JM, Canizares J, Nuez F. 2009. Transcriptomic analysis of tomato carpel development reveals alterations in ethylene and gibberellin synthesis during pat3/pat4 parthenocarpic fruit set. BMC Plant Biol 9:67.

Peddada SD, Lobenhofer EK, Li L, Afshari CA, Weinberg CR, Umbach DM. 2003. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. Bioinformatics 19(7):834-41.

Pemov A, Park C, Reilly K, Stewart D. 2010. Evidence of perturbations of cell cycle and DNA repair pathways as a consequence of human and murine NF1-haploinsufficiency. BMC Genomics 11(1):194.

Perrin B-E, Ralaivola L, Mazurie A, Bottani S, Mallet J, d'Alche-Buc F. 2003. Gene networks inference using dynamic Bayesian networks. Bioinformatics 19(suppl_2):ii138-148.

Prickett D, Watson M. 2009. Use of GenMAPP and MAPPFinder to analyse pathways involved in chickens infected with the protozoan parasite Eimeria. BMC Proceedings 3(Suppl 4):S7.

Qian Y, Khattak SF, Xing Z, He A, Kayne PS, Qian N-X, Pan S-H, Li ZJ. 2011. Cell culture and gene transcription effects of copper sulfate on Chinese hamster ovary cells. Biotechnology Progress 27(4):1190-1194.

Quackenbush J. 2002. Microarray data normalization and transformation. Nat Genet.

Quinlan JR. 1993. C4.5: Programs for Machine Learning: Morgan Kaufmann Publishers.

Ramaker H-J, van Sprang ENM, Westerhuis JA, Smilde AK. 2003. Dynamic time warping of spectroscopic BATCH data. Analytica Chimica Acta 498(1-2):133-153.

Ramoni MF, Sebastiani P, Kohane IS. 2002. Cluster analysis of gene expression dynamics. Proc Natl Acad Sci U S A 99(14):9121-6.

Richards EJ, Elgin SCR. 2002. Epigenetic Codes for Heterochromatin Formation and Silencing: Rounding up the Usual Suspects. Cell 108(4):489-500.

Ronald Zielke H, Ozand PT, Tyson Tildon J, Sevdalian DA, Cornblath M. 1978. Reciprocal regulation of glucose and glutamine utilization by cultured human diploid fibroblasts. Journal of Cellular Physiology 95(1):41-48.

Sakoe H, Chiba S. 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 26(1):43-49.

Sanges R, Cordero F, Calogero RA. 2007. oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language. Bioinformatics 23(24):3406-8.

Sauerwald TM, Figueroa B, Hardwick JM, Oyler GA, Betenbaugh MJ. 2006. Combining caspase and mitochondrial dysfunction inhibitors of apoptosis to limit cell death in mammalian cell cultures. Biotechnology and Bioengineering 94(2):362-372.

Schachtner R, Lutter D, Stadlthanner K, Lang EW, Schmitz G, Tome AM, Gomez Vilda P. Routes to identify marker genes for microarray classification; 2007 22-26 Aug. 2007. p 4617-4620.

Schaub J, Clemens C, Schorn P, Hildebrandt T, Rust W, Mennerich D, Kaufmann H, Schulz TW. 2010. CHO gene expression profiling in biopharmaceutical process analysis and design. Biotechnology and Bioengineering 105(2):431-438.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science 270(5235):467-470.

Schliep A, Schonhuth A, Steinhoff C. 2003. Using hidden Markov models to analyze gene expression time course data. Bioinformatics 19 Suppl 1:i255-63.

Schwammle V, Jensen ONJ. 2010. A simple and fast method to determine the parameters for fuzzy câ€“means cluster analysis. Bioinformatics 26(22):2841-2848.

Secco M, Moreira Y, Zucconi E, Vieira N, Jazedje T, Muotri A, Okamoto O, Verjovski-Almeida S, Zatz M. 2009. Gene Expression Profile of Mesenchymal Stem Cells from Paired Umbilical Cord Units: Cord is Different from Blood. Stem Cell Reviews and Reports 5(4):387-401.

Seth G, Charaniya S, Wlaschin K, Hu W. 2007a. In pursuit of a super producer—alternative paths to high producing recombinant mammalian cells. Current Opinion in Biotechnology 18(6):557-564.

Seth G, Charaniya S, Wlaschin KF, Hu W-S. 2007b. In pursuit of a super producer--alternative paths to high producing recombinant mammalian cells. Current Opinion in Biotechnology 18(6):557-564.

Seth G, Philp RJ, Lau A, Jiun KY, Yap M, Hu W-S. 2007c. Molecular portrait of high productivity in recombinant NS0 cells. Biotechnology and Bioengineering 97(4):933-951.

Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, Shiloh Y, Elkon R. 2005. EXPANDER - an integrative program suite for microarray data analysis. BMC Bioinformatics 6(1):232.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498-504.

Sherlock G. 2000. Analysis of large-scale gene expression data. Current Opinion in Immunology 12(2):201-205.

Singh A, Elvitigala T, Cameron J, Ghosh B, Bhattacharyya-Pakrasi M, Pakrasi H. 2010. Integrative analysis of large scale expression profiles reveals core transcriptional response and coordination between multiple cellular processes in a cyanobacterium. BMC Systems Biology 4(1):105.

Smilde AK, Hoefsloot HCJ, Westerhuis JA. 2008. The geometry of ASCA. Journal of Chemometrics 22(8):464-471.

Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers R-JAN, van der Greef J, Timmerman ME. 2005. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics 21(13):3043-3048.

Smith AA, Vollrath A, Bradfield CA, Craven M. 2009. Clustered alignments of gene-expression time series data. Bioinformatics 25(12):i119-1127.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology 3.

Spens E, Häggström L. 2009. Proliferation of NS0 cells in protein-free medium: The role of cell-derived proteins, known growth factors and cellular receptors. Journal of Biotechnology 141(3-4):123-129.

Storey JD, Dai JY, Leek JT. 2007. The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. Biostat 8(2):414-432.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 100(16):9440-9445.

Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. 2005. Significance analysis of time course microarray experiments. Proceedings of the National Academy of Sciences of the United States of America 102(36):12837-12842.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and others. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102(43):15545-15550.

Suzuki H, Forrest ARR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T, Ravasi T, Hasegawa Y, de Hoon MJL and others. 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nature Genetics 41(5):553-562.

Swiderek H, Logan A, Al-Rubeai M. 2008. Cellular and transcriptomic analysis of NS0 cell response during exposure to hypoxia. Journal of Biotechnology 134(1-2):103-111.

Tai YC, Speed TP. 2006. A multivariate empirical Bayes statistic for replicated microarray time couse data. Annals of Statistics 34(5):6.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences of the United States of America 96(6):2907-2912.

Tan P-N, Steinbach M, Kumar V. 2005. Introduction to Data Mining: Addison-Wesley.

Tarraga J, Medina I, Carbonell J, Huerta-Cepas J, Minguez P, Alloza E, Al-Shahrour F, Vegas-Azcarate S, Goetz S, Escobar P and others. 2008. GEPAS, a web-based tool for microarray data analysis and interpretation. Nucleic Acids Res 36(Web Server issue):W308-14.

Taylor RC, Acquaah-Mensah G, Singhal M, Malhotra D, Biswal S. 2008. Network inference algorithms elucidate Nrf2 regulation of mouse lung oxidative stress. PLoS Comput Biol 4(8):e1000166.

Taylor RC, Singhal M, Weller J, Khoshnevis S, Shi L, McDermott J. 2009. A network inference workflow applied to virulence-related processes in Salmonella typhimurium. Ann N Y Acad Sci 1158:143-58.

Tchagang A, Bui K, McGinnis T, Benos P. 2009. Extracting biologically significant patterns from short time series gene expression data. BMC Bioinformatics 10(1):255.

The ENCODE Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.pdf. Nature 447:799-816.

The FANTOM Consortium. 2005. The Transcriptional Landscape of the Mammalian Genome. Science 309(5740):1559-1563.

Tong W, Hong H, Fang H, Xie Q, Perkins R. 2003. Decision Forest: Combining the Predictions of Multiple Independent Decision Tree Models. Journal of Chemical Information and Computer Sciences 43(2):525-531.

Tusher VG, Tibshirani R, Chu G. 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(9):5116-5121.

Ulloa-Montoya F, Kidder B, Pauwelyn K, Chase L, Luttun A, Crabbe A, Geraerts M, Sharov A, Piao Y, Ko M and others. 2007. Comparative transcriptome analysis of embryonic and adult stem cells with extended and limited differentiation capacity. Genome Biology 8(8):R163.

Ündey C. 2004. Intelligent real-time performance monitoring and quality prediction for batch/fed-batch cultivations. Journal of Biotechnology 108(1):61-77.

Ündey C, Ertunç S, Mistretta T, Looze B. 2010. Applied advanced process analytics in biopharmaceutical manufacturing: Challenges and prospects in real-time monitoring and control. Journal of Process Control 20(9):1009-1018.
Vega F, Coombes KR, Thomazy VA, Patel K, Lang W, Jones D. 2006. Tissue-Specific Function of Lymph Node Fibroblastic Reticulum Cells. Pathobiology 73(2):71-81.
Vis D, Westerhuis J, Smilde A, van der Greef J. 2007. Statistical validation of megavariate effects in ASCA. BMC Bioinformatics 8(1):322.
Wang J, Bø T, Jonassen I, Myklebost O, Hovig E. 2003. Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. BMC Bioinformatics 4(1):1-12.
Wang J, Reijmers T, Chen L, Van Der Heijden R, Wang M, Peng S, Hankemeier T, Xu G, Van Der Greef J. 2009a. Systems toxicology study of doxorubicin on rats using ultra performance liquid chromatography coupled with mass spectrometry based metabolomics. Metabolomics 5(4):407-418.
Wang X, Wu M, Li Z, Chan C. 2008. Short time-series microarray analysis: methods and challenges. BMC Syst Biol 2:58.
Wang Z, Gerstein M, Snyder M. 2009b. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57-63.
Way KJ, Dinh H, Keene MR, White KE, Clanchy FIL, Lusby P, Roiniotis J, Cook AD, Cassady AI, Curtis DJ and others. 2009. The generation and properties of human macrophage populations from hemopoietic stem cells. Journal of Leukocyte Biology 85(5):766-778.
White P, Lee May C, Lamounier RN, Brestelli JE, Kaestner KH. 2008. Defining Pancreatic Endocrine Precursors and Their Descendants. Diabetes 57(3):654-668.
Wilson C, Bellen HJ, Gehring WJ. 1990. Position Effects on Eukaryotic Gene Expression. Annual Review of Cell Biology 6(1):679-714.
Wirth M, Bode J, Zettlmeissl G, Hauser H. 1988. Isolation of overproducing recombinant mammalian cell lines by a fast and simple selection procedure. Gene 73(2):419-426.
Wlaschin K, Hu W. 2007a. Engineering cell metabolism for high-density cell culture via manipulation of sugar transport. Journal of Biotechnology 131(2):168-176.
Wlaschin KF, Hu W-S. 2007b. Engineering cell metabolism for high-density cell culture via manipulation of sugar transport. Journal of Biotechnology 131(2):168-176.
Wong CE, Singh MB, Bhalla PL. 2009a. Floral initiation process at the soybean shoot apical meristem may involve multiple hormonal pathways. Plant Signal Behav 4(7):648-51.
Wong CE, Singh MB, Bhalla PL. 2009b. Molecular processes underlying the floral transition in the soybean shoot apical meristem. Plant J 57(5):832-45.
Wu C-J, Kasif S. 2005. GEMS: a web server for biclustering analysis of expression data. Nucleic Acids Research 33(suppl 2):W596-W599.
Wurm FM. 2004. Production of recombinant protein therapeutics in cultivated mammalian cells. Nature Biotechnology 22(11):1393-1398.
Wurm FM, Gwinn KA, Kingston RE. 1986. Inducible overproduction of the mouse c-myc protein in mammalian cells. Proceedings of the National Academy of Sciences 83(15):5414-5418.
Wurm FM, Petropoulos CJ. 1994. Plasmid Integration, Amplification and Cytogenetics in CHO Cells: Questions and Comments. Biologicals 22(2):95-102.
Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, Chen W, Xie M, Wang W, Hammond S and others. 2011. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nat Biotech 29(8):735-741.
Yee JC, de Leon Gatti M, Philp RJ, Yap M, Hu W-S. 2008. Genomic and proteomic exploration of CHO and hybridoma cells under sodium butyrate treatment. Biotechnology and Bioengineering 99(5):1186-1204.

Yee JC, Gerdtzen ZP, Hu W-S. 2009. Comparative transcriptome analysis to unveil genes affecting recombinant protein productivity in mammalian cells. Biotechnology and Bioengineering 102(1):246-263.

Yeung KY, Ruzzo WL. 2001. Principal component analysis for clustering gene expression data. Bioinformatics 17(9):763-774.

Yoshikawa T, Nakanishi F, Ogura Y, Oi D, Omasa T, Katakura Y, Kishimoto M, Suga K-i. 2000. Amplified gene location in chromosomal DNA affected recombinant protein production and stability of amplified genes. Biotechnol. Prog. 16(Copyright (C) 2012 American Chemical Society (ACS). All Rights Reserved.):710-715.

Yu X, Griffith WC, Hanspers K, Dillman JF, Ong H, Vredevoogd MA, Faustman EM. 2006. A System-Based Approach to Interpret Dose- and Time-Dependent Microarray Data: Quantitative Integration of Gene Ontology Analysis for Risk Assessment. Toxicological Sciences 92(2):560-577.

Zhan M, Yamaza H, Sun Y, Sinclair J, Li H, Zou S. 2007. Temporal and spatial transcriptional profiles of aging in Drosophila melanogaster. Genome Research 17(8):1236-1243.

Zheng H, Ying H, Yan H, Kimmelman AC, Hiller DJ, Chen A-J, Perry SR, Tonon G, Chu GC, Ding Z and others. 2008. p53 and Pten control neural and glioma stem/progenitor cell renewal and differentiation. Nature 455(7216):1129-1133.

Zhou M, Crawford Y, Ng D, Tung J, Pynn AFJ, Meier A, Yuk IH, Vijayasankaran N, Leach K, Joly J and others. 2011. Decreasing lactate level and increasing antibody production in Chinese Hamster Ovary cells (CHO) by reducing the expression of lactate dehydrogenase and pyruvate dehydrogenase kinases. Journal of Biotechnology 153(1-2):27-34.

Zhou W, Chen C-C, Buckland B, Aunins J. 1997. Fed-batch culture of recombinant NS0 myeloma cells with high monoclonal antibody production. Biotechnology and Bioengineering 55(5):783-792.

Zhu J, Chen Y, Leonardson AS, Wang K, Lamb JR, Emilsson V, Schadt EE. 2010. Characterizing Dynamic Changes in the Human Blood Transcriptional Network. PLoS Comput Biol 6(2):e1000671.

Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE. 2008. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet 40(7):854-861.

Zoppoli P, Morganella S, Ceccarelli M. 2010. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. BMC Bioinformatics 11(1):154.

Zou C, Feng J. 2009. Granger causality vs. dynamic Bayesian network inference: a comparative study. BMC Bioinformatics 10(1):122.

Zou M, Conzen SD. 2005. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21(1):71-79.

9 APPENDIX

The following sections provide the sequences of the three genes from which promoters were isolated as described in Chapter 6. These sequences were derived from the Chinese hamster genome scaffolds, which were assembled in our laboratory in March 2011. Briefly, genomic DNA from the liver of a highly inbred 17A/GY female Chinese hamster was used for sequencing. Over 275 giga base pairs of sequences were generated using the Illumina paired-end technology. About 10% of the reads of low quality, homo-polymers, and short tandem repeats, were removed. The remaining sequences were assembled *de novo* using Assembly by short sequences (ABySS), a parallelized sequence assembler. The assembled contiguous sequences (contigs) were further linked using the SSPACE scaffolder. About 32,590 scaffolds with a minimum length of 1 Kbp and an N50 of 2.2 Mbp were obtained. The estimated size of this draft version of the Chinese genome is approximately 2.52 Gbp. More than 32,000 putative genes were predicted in these scaffolds using *ab initio* prediction by Augustus and BLASTing against our annotated transcript sequences. More details about the sequencing, assembly, and analysis of the Chinese hamster genome can be found in (Jacob et al.).

The sequence formats are as follows:

- Red letters: untranslated regions (UTRs), identified by BLASTing against the mouse genome, version mm9 (http://genome.ucsc.edu/)
- Blue letters: exons, identified by BLASTing against the mouse genome, version mm9 (http://genome.ucsc.edu/)
- Underscored letters: supported by our transcript sequences (S5.1)
- Highlighted in grey: supported when genomic reads were used to extend exonic regions

9.1 TXNIP GENE SEQUENCE IN CHINESE HAMSTER GENOME

1622000..1629000, scaffold 122

GTCAACACCCTGGGACCCGGTAACAGCCACCTCTCCTCCTGTCTTCTGATCTCCTCAAA
TCTTAAACAACAACAACAATAATTGTTTACATTTCACTTATTTTGTGTTCTCTGGGAGT
GAGCAGTCGGGTGGTCGGGGAGGTCAGGGTTGGTTCTCTCCTACTTGGGGGATTCTGGG

GGTTGAACATACCACACTAGTTAACAATATCGATATATACACTTTTTGGTAAACATTTT
GATGTTTTATAACAACAAGTTATTTTTTTTTCCAAGATGCCTACTACAGTTTGGAACTA
AAATTAGTGAGAGTCCCACCTCCTTTGTTCTTAATCAATACTGTTAACTCAAGCCTAGT
TACCTTTCTGGAACAAACACACACCCTAAGGATCAACATGTTCCTAGGCTTTGGTGT
TGTGTGTATGTTTTTCAATCCTGTGATGTCTAACTCTGAGCTTGCTTAACATTGAGGTT
GAGTTGCAACTTGTTAGCCTTAGGATGTCAACCCAGTTAAGGCATGTGAGCAAGGGGAG
GGTGCTGTTAAGAGCAGTGAACAGGTTACTAATTGTGCAGAAGGCAGAATTTCATATAT
GCAAAGGCTTTTTGCTTCCCCATCTTAAATTCAGGACCATTGTCTTCATTTTCCAGCAC
AAGGAACAAAACAACAATCTCCTAGTAACTCTGCCTAAATATGGAAGGAGTCTAGCATT
CAAAGACAAGGCGGGGCGGGGGTGGGGGGGTGATCCCTCTTCTACACAAACATTTTCTT
AGCACATTCAAGAGTTATTTCTTGAGTCAGCACCTGGAGGCACCTCTGTCAGTTTGGTT
TACCAGATAGCCAAGTGAAGCCAATACAACCTGGGCAACAGGCTTAGGGATATGCATTT
CATTTAGTGAGTCTGGAGAGATGATGAAGACAAAAAAAAAAAGTGACAGGAACTCTGGA
ACAAAGTCAGGAGTGATGGATCTTTTCCTTCTTTCCTTCCTTTATTTATTTATTGTACC
CACCCCTTGTTTCCTGGAGAAACAGGGACGGAGGGAGACCTCAGTGTGAAGTACACACC
TCATTAAAGCAGCTATGAGGTAAATGAGGGAACATATACAAAGTGTTCCCCAATTTCAC
AGGTAGACTGAAGAGACTTGGGGAAAAGCAACAGGATGTGGACACTCCCTTACTGCTTC
CGTTCTAGAGAACAGAGAACAGAAGAGAGTGCACACGGGTGAGGGACAGTCCATAGGGG
CGTGAGGGGAAGGGGGCACCCCAGAACAAGGTCTCGGCTTTCCCATGAGGACTTCTGAA
AGGGGGCTTAAAATATCTGACACACGCGTTTTGTTCTTCAGTAGACATGCAACGGGAAG
ACACCGGGTGGGCTTTCTCTAACACAGCCGCACCCTCACCCTCCACCCAATGCTGAAGA
ACCCTTGGGGCAATGGGATTGTTTTTCACAGACTTGGCGAATTTGGAAGCCAGGAATAA
ACAGTCGCCTCGAAATGAATTGCGCTGGCTAAGAGGGGGCATGAAATCCTCTCCTAAGC
ACATTTTCCTTCTCCTAAAAAAAGAAAAAAAAAAAAATCACACNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCACGCGCCCCTGCGGCC
GCGCGGATTGGTCGGAGGCCTGGTAAACAAGGGCCAGGCAGCCAATGGGAGGACTGTGC
ACGAGGGCTGCACGAGCCTCTGAGCCAGGGCTCGCGTGGGGCTCCTAGCCCGGCAGCTA
TATAAAGGCGTTTCCGGCTCCCGATTGACACTCGCCACCCCTCATCTCGTGGGTTTCCT
GGGCTCATCCGGCTGCTGGAGCAAAGAGACAGCCGCCGCTGCTTTCCTCTTGAAAGCCT
TGGGCACAGCCAGCAGGGTCGAAGCTCTCAGGCACCGTTTCGAGAAGTTGTTGGGGTTT
TGTTAAAGGGCTCCTTTAGCTTTTCGTTTGGGTTTTTTCAAGCCCTGTCTGCCTTCCGG
AGAGTGAACTCTTCATCTGCTGTGAAGGGCTCTTACGAGTTTTCTCTCCCGCCTCCGTT
TATCTGGAACCCGCTGAACCCAATCATGGTGATGTTCAAGAAGATCAAGTCTTTTGAGG
TGGTCTTCAACGACCCCGAGAAGGTGTACGGCAGCGGGGAGAAGGTGGCCGGCCGGGTG
ATAGTGGAAGTGTGTGAAGTTACTCGGATCAAAGCCGTCAGGATCCTGGCTTGCGGAGT
GGCCAAGGTCCTGTGGATGCAAGGGTCTCAGCAGTGCAAACAGACCTTGGACTACCTGC
GTTATGAAGACACGCTCCTCCTAGAAGACCAGCCAACAGGTACAGACACAGATAGCTCC
AGCTGTTCTTCCTCGCACGGGCGCTTTGACAAGGGGAGGCACGGGAGCCTCCAGCAAGG
GTTTGGAGTTGAGTCGTTCGTTTAATAGTTTGAATTGTGTTGGTTCCGATACTTTTAAG
TTTTAATTTGCAAGTATTTCACCAACCCCTCCCCCACCCCCGCAGGGGTGATTTGTAGT
TTCTGAGGTGCTGTTTGCTAAGGTGAAGCCTTAGCTACTTAAAGATTAATGTATTCGCG
GTCTGTACCCGATTCCTGGGGCAACCGTGCTGGATGGGCCTTTTCGTAGTTGTGGGCTT

GGCTCTGCAACTTTCTGTCCAAGAAAGTGCACTGGAATCTTTTATTCCACCGTGCAAGA
GATGGACGTGGAAGTAAAATAACCTTGTCTGCTGCCCAGTTTTCACAGATGCCTTTGAA
ATGTAATAGGGAACCTAATGGACCCTTGCCTTTGTTTCCCCGCAGGTGAGAACGAGATG
GTGATCATGAGACCTGGAAACAAATACGAGTACAAGTTTGGCTTTGAACTTCCTCAAGG
GTAGGCATCTACCAAATGCACCTTGGACTCTGTTTCTAAAAGCCCCACCCCCATTGACC
TCTTACTGTTCTTAGAGAGCATTAATTTTTTTTTCATTGTTTTGTCACAGGCCCCTGGG
AACATCTTTCAAAGGAAAATACGGTTGTGTCGACTACTGGGTGAAGGCTTTCCTTGATC
GCCCCAGCCAGCCGACTCAGGAGGCAAAGAAAAACTTTGAAGTGATGGATCTAGTGGAT
GTCAATACCCCTGATTTAATGGTGAGATTTAGTTCTCCTTGTTTGGGGATAACAAATTA
GATGCTTGGGGCATGGAAATAACTCAAAACATTGTGTCCTCCTACACAGGCACCGGTAT
CCGCTAAAAAGGAGAAGAAAGTTTCCTGCATGTTCATCCCTGATGGGCGTGTGTCCGTC
TCTGCTCGAATTGAAAGGAAAGGATTCTGTGAAGGTAATATCCTCATGCTTAAAATGTA
GCCAGGGTGGCCAGAGCCGTGGGCTTGGAGGGGGGCATGGAGGAGAGGAAGTGCCGTTA
AGTAAACGGATGTTCATCCCTCTTCATCTTGAATCCCAGGTGATGACATCTCCATCCAT
GCCGACTTCGAGAACACATGTTCCCGGATTGTGGTCCCCAAAGCAGCTATTGTGGCTCG
GCACACTTACCTTGCCAATGGCCAGACCAAAGTGTTAACGCAGAAGCTATCCTCAGTCA
GAGGCAATCACATCATCTCGGGGACCTGCGCCTCGTGGCGAGGCAAGAGCCTCCGTGTG
CAGAAGATCAGGCCATCCATCCTGGGCTGCCACATCCTGCGCGTGGAGTATTCCCTGCT
GGTGAGTGTCATGGCGGCTTGGTCCTAGAAGAGAAACAATTCTTGTGTTACAAATTGAG
TGCTTTCTCTACACAACCCTTCTGACAAACTGCCACCTTGCTTTACAGATCTATGTTAG
TGTTCCTGGCTCCAAGAAAGTCATCCTTGACCTTCCCCTGGTGATTGGCAGCAGGTCAG
GTCTGAGCAGCAGGACTTCCAGCATGGCCAGCCGAACGAGCTCTGAAATGAGCTGGATA
GATCTAAACATCCCAGATACCCCAGAAGGTGGGTATCTTGCTTTCTTTGGGCCTGACGG
GGCTTTGTGAAATTGTGGTGATCTGGCAGGTTTTGGCTGACTTTGTCATCCATTTTCTT
TGTAGCCCCTCCTTGCTATATGGACATCATTCCTGAAGACCACCGACTAGAGAGCCCCA
CCACGCCTCTGCTAGATGACATGGACGGTGCTCAGGACAGCCCTATCTTTATGTATGCC
CCCGAGTTCCAGTTCATGCCTCCACCCACCTACACTGAGGTGAGATGTGCGATTTTTTT
TTTTTTTTTTTGACAGTTTGTCTCAGTTTGTCCTGAAGGAAGAAAGAGTTATTAAACTC
ATGGCTTTTCTACTACTTAACCTAAAAAACTAAAATGCTGTTTTCCCTCTTCCTCCCCC
AGGTGGATCCCTGCACTGTTAACAATAACAACAACAACAACAACAACAACAACGTGCAG
TGAGCCCGTGGGAGAAGAGATGCGTCTATACTCACTCATTTCTTTCCCCTCTCTGCTTG
GACGCCAGTGTTTCAGAGACTTAGTCTGAAAGTGGAGAACGGGTTCACCCCAGCCTCTG
ACTCCACATCTGGGTGATCAACAGGCGGGTCCCTGGCTTCAAGTGGCGCAGACACAGCC
AGTGCCCGCCCTGTGGTGTAGGAGCGTTTGCTGGGTGGATAGAAGAACACTCTAAAAAA
TTCAGACCCCGTCCACTTTCTCTCAGATCTTGGAAATGAAGACATTGTCCGCAGTTTTT
GAGTCGGTGTGAATGACCTTCTGGCATGGTCTTTACAAGGTTTTTTTTTATTTGGAGGA
GTTACAGGTTAACAGCAGACCATGCATTTCTGTGCCATGGGGGACAGAATCAGCTTCAC
ATACTAGATAATCATGGCCAAAGTTTGGAAGAGGTGTTTTTTTATTAGTAATATTTTTT
TAATCAGTGTTCCCTTTTTATAACCTTAAAAGGAAAATGAAAATCTTCCAAGGCTGTGT
GGTTTAGGCTTGGGTGATGGTGGCCACCCTGTGACTTACAGGGTAGCTGCTCCCAGTGA
ACTTTAACCTAGGCTCAGAGCAGTCGTGGCTGCACCCACCAACAGCCATAAAAGCCATT
TTACAGCCAGTTGCACTGTGTTCTCTTACAAACTTAATCAAATGGGAGAATCTGTTATT
TCCGTGTGACTCCTTGGAATTGATTCTAAGGTGATGTTCTTAGCACTTTAGCTCCTGTC
AATTTTGTTTTAATCTCGATTGCAGATGTAAACTCTACTACTATGTGTCTTAAGGGTTA
AGCCCAATGACAGGGCAAAATGGATTTTTGTGTTTGGGTTTTTTTTTTTTTTTTTGAGA
TGGGGTCTCTCTATCCTTTGTTAGCCTGGAACTCTACATAGACCAGGCTGGCCTCGGAG

ATCCCTCTGTCTCTGCCTCCTGAGCGCTGAGATTGAAGGCTCATCTTGGTTATTACTTT
TTGAGACTTTGTCCCAGTGTGCAAAATTGTGTAATTCTGATTCGAATATGAAGGGTTGC
CGCCCTGTGTGTAAGCCTCTGCATTGAGATACCGCAGCGCTGATCAGGGTCTCCATGCG
TTGCAATGAGAACTGCAGGTTTTCACGAAAGCCCCCAGACCCATGTTCTGCCCAGTGTT
AGGAGAAAGGACTCTTAACTGCTTTCACTGTGGCCACTAGCACCTGATCAGTCACTTTC
AGCCATAGCACTTTGTCCATTGTCCTGTGTCAGCACTGAGTTCCACCCTTTTCTGAGAA
CTATGCAATCAGAAAGCATGTGAGCTGATGGTGGTTTCATTTTTGTTGTTTTTGTATTT
TGTGTATCCTTTTTGTATGACAAAAACTATATTTTGTACTTATCCAAATATATTTTCAC
CCCAAGGTGGGGATATCCTTTGTAAAAAAATAAAATAAAAATAAAGTTTTTTAATGGCA
AAAATGTTTTTTTTTCTTGTCTGAAATATTAGAGAAGTTACTAGAAAACAGGTGTGCAT
ACTAGTTTTCCCCTGCGTCACAATGCTGAAAACCCTGTGGAGAGAGCAGTCCCTATAGA
CCCTAAGTGACACTGGAGTTGAAGGGGAACAGCAGTCCGTCAGCTGTTTTGGCCTCCAT
CAAGGCTACATTGGTGGCATCTGGTGATGAAAGCTAATTAACCACATCCTGAAAATTCT
TCAGACCTCCTCCGGTAGCTAAGGATTGGTTCAAGGTGCTGGGATCAAAGGCATATATT
ACCATACCCAGGTTTTTCTTTTAGAGCCTGCCTTGCAAGAGTGTGGGACCCGGAGACCA
GGTCCTCGGGGTTTAAAGAAAGTCTACCAGCTGAGCCTTCTCAACATACAGTTCTCCAT
GCTCACTCAGTTTTATGAATTCTTGAAGAAATAGCCTTGAGGGGATTTGGAAAGGAAAG
AGTGAGTTTCCTTTGTGTATTTATCTTTTAAACTGTGTGAGCAGCTTTTATTTACACAA
AGGCGGCAGGAGGATCTGGGGAGAAGGGACTGGAGACTAACGTGAGGGAAATTAAATAG
GTGTGACTCGGGTGGGAGGAAATCAATACTGGAGCCCAATGGATCCCCTTATGGTCCGT
AATGTTCATCGGACACTGTAAATACAGAATGAGAATGGGTCAGGACTGGTAGAGGCCTC
CATCCACCTACCCCAAGGATGGTTTGTGGGGATGGAGTTTTGGTTTTGTTTTGTTGAGA
CAAGTTCTCTGTAGCTCAGGCTAGCCTCAGCCTCCCAGACACTGGGGTCAGAATCATAT
ACTACCATTCCTGACTGACATTCTTTCTTTCTTTCTTTC

9.2 MMP12 GENE SEQUENCE IN CHINESE HAMSTER GENOME

1272000..1288500, scaffold 8

ATCCATTATTGCAGGAAATTCAAGCAATTGAAGTTTTCTTCCTGCATGAGCAAGCCAGT
GGTGTGTGAAAGGTCAGTGAAAGCATGCATTTCTTTGTGCATTTTTGGAGCATAGCTCC
TTCCTGGTCCTTCCTGGAAACCAGTTTGCCTCTGGATTCCCGGGGAATGGCACTACAAA
TGGTTTTGGGTATCATTGATACACTCAACTGTCCTGATGTCTGTCTGAGGATGCAGCAA
TGTCTAAATTCTTGTACTTTTGGAGACTTTATGTTATAAAATATGCCTTCTACTTAAGG
AAATTCTGATGAGTACCCATCATTTTGTGTATGAAGAACTCAGGCAGTGGCTGGTTTCT
GCATTTCCATAAAGGTGCTGTTTGTGTTGAGCATCATTGAAGGGCATTTGATTCTCAAT
CCCCATGGTGGCTCTTCTGGGATCCTCAGCTTTAAATCAGGACCTGGGTCAGGAGCTTC
CTGTAGTTCTTTCTGTAAAGAGAGCTGAGCAGTAGTCAATCTTCCTCTTCATTCTCATT
ACCAACTTAAAAAAAAAATTACCTTACATGGGAGTGCAGAAAGAAAGCATGAAGGAATA
TTGCAATCATGGAATTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTAAA
AGAGAGAGAGAGAGAGAGAGAGAGAGAGNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGAGAGAGAGAGAGAGAGAGAGA
GAGATTTTACTGAGTCCATTTAGCATACAGTTTTGCTTCTGTGTGTATGTTTTAGGGGG
CTGTTTTGTTTTGTTTTTAGAGACTGGGTTCTCTGTGTAGCTTTTGCAGTCAGTCCTGG

AACTCATTCTGTAGACCAGGCTGGCCTCAAACCCATGGAGATCTGCCTGCCTCTGCCTC
CTGAGTGCTAGGATTAAAGTTGTGCTCCACCATAACCCAGCCTCTGTATGTTTCTAAGG
CTGATCACTTGGAGTTGGATAAGATATCAAGGGGGCTTATCCGCTTAGCCCCAGAAAAA
AACTGATTCTTTTTCTCTCAGCCACCATTGACTGCTTGTAGCTTTCATTCAGGAGCCTT
GTAATATGCACCCCATCCATGTTGACAAGTTAACTGGTATTGCAATTACACAGGCCTTG
CTTAAGTGAGCGTATTGTTGAATTTTCATGGGTGCAACATCCCTGTCATATCTAGAAGA
CACTATCTCACATCAGGGTCCTGGACCTTTGTTGCTTGTAATCCTTTCACCACCTCTTC
TCTGATTTTCTCTGAGCTGTACGTGTAGGGTTGCATTGTAGGTGACCTTTAAGCTGTGC
CAGGCCATTGGCTCACTACTTTTCCCTATTATTTATAACCATCCCTGGTGTAGAAACAA
TTTTTACTTCAACTTTACAAGTGGAAAACTCATCTTGAAGGTCATAAAACACCAGACTA
AGATCCTTAATCAGGTCAGCTTTAGTCTAGGAATTAACTTGACAACCTCGGAAGTGCCT
GTAGCTACCCATGTAGTCTTTCTCTCCTTACACTTGACTAATTTGGAGCAACTGGTGAG
TTGAGACACAAGAGAACCAAAACGTGTTCACCTACTCTCATTTCTGTTTAGGGCACTCT
CACTTCTTAGTATCTCTGTTATCCTCTTGCGTCTTGCATCCTACATGCACGAGAAATTT
CCTGCAAATCTTCCAAGAACTGAATTCTGTCTTGCTGTGGATGAGACCAGTTCTTGCTT
TCAGTGGATCAGGAACACACCCTCCTGGTAGTGATTGCTCCCACCTGTGAGGTTCATTT
CCAATGTCCTTAATGTTCTCCAACCAGACTCAGTGGTACTGAGTGTCTACTACCCTCAA
GGTACTTTTCAGCACCAGGAATACACAACAAACAGACAAACAAAAAACCTCCATTTTCC
CACTTTGTAGCCCAACCCAGTTCTGTTCCTCTTACGATTATTGAGTTAGATTCGGTATT
TGAAAGGATTTCCTTACAATGTTCAACACATGATCATTCTCTTTCTGAACCCTTGAGCT
TCTTCTGCTTTCCAATTCAATGCTTGATAATCCCTGTTAAGGACAAAGGGAATTGACTG
GGTAACAGTGGGGAACTTCCTGTCACTAGAGTGCTTAACTCACTGTCTCATTTAATTGT
CACAATAGCATAAGTAGGTTAAAGCATATTGTCCCCATTCCAAGTGGGTCACCTAGAAA
TCAAGCCACAGAATATGAAAGAGGTAGGGTAAGGATACGGGATAGTGTGGTGGCAGGGG
TGGGAGTGTGTGTTCCTTCTCTTCCTGTCTGAAACTTCTAAAATCGTTAGAAAAGGGGA
ACAAAGGAGAACAAAGCATAGGACACATGTGTATGAAGAAGCCATACTGGATCTGTTGT
TTTGCATGATAACCGAAACATCAATATTCATTCATTTGTTTATTATATATGTCTCTCTG
TGTGCATGCATGCCATAATAGGCTTGTGAAGGCTAGCGGACAACTTCCAAACATGTGTC
CTCACCTTCCACTCACTATGTGGGTTCCTGGAGATCCAACTCAGGTCATGAAGTTTTGG
AGGAAGTATCTTTGTCTGCTAAATCATCTAATAGGCCCCTAAAATTCATTAAAATTAAC
ATTAGGAAGTAGCCAGCATGTCCCATCTCAGAGAGATTGGAGAGAGTAGCCCACTTGTA
CACCCATACCCTAGACTATTCATCTAGAAGATGCTTGTGATGGGATTAAAGGGGTTAGA
CAGAGTTCTGTCCACACAGCATATTCTAGCAAAAGTATTTTCTCCACACATTCTTTTAA
AATCTCCTTTAAAGCACTTCTTCTGTAATGAGTATAAGGAAAGCATTGTAAGGAGAATG
GACAACACCAAGCTTTCCAAAACTCTACTGGCAGAGTTTTGGTAGATTTGTTGTTGTGT
CTTACCTATGCTTATAAGTAAGTAGAAGCTCAGGCTCACTGGTCATTTTGAAGAAATAC
TCTAGACATGATTAGGTGAGTGATGACTTGTATGGCACTTTGACCCACTGTCCATTCAC
ATGCTTCTGAAGTACAGTAAGAGATGATATCAACTGTGAGTCACTCAGAGGATTTCACG
TTCACAGATCCGGACTAAAGTGTTATATAAAGCGGACTTTAAGGGAACTTAGAGAGCAG
AGCTCTGCTGAGAGATCTGGTCACAATGAAGTTCCTCATGCTCCTCGTGGTCTTGCAGG
TGTCTGCCTGTGGGGCCATTCCTATAAATGAAACCGAATTTGCTGAAGTGAGTATGACA
GGGCTGAACTGGCCTAGCCTGTGTCTTTAATATCTACTGAATTTGTAAGTTTTATTTTG
AACAGTATGATTTGATTGGAAAAATCAGCATAGGTAGCTGTTTTCTTTTCAGCATGTTT
GGTTTTCCATCGAGGAATATCCAAATGAAGTTATGATTTTGAAAGAATGTTATTAAATT
GGTGAATAATTGTGCTATCTACAAAAACCTATGGGTAAATAGTTTAGCTCAATATAATT
GAATTACATGTCTAATAGAGTGCTGTCTTTTGGTCTCCCACCCCCATTTTGTCTTTTAT

TTGTATTAGTAATTTAATGTCCATTAATAATGTGAGATTTGGACAAGGTGTAATTGATG
TAATAGCTTCAAATTTTCATGATAGTTACAGCATTCTCCCTCCATGTTCATTTACCTCA
GATATTTATTCATTTCATAAGTGTTTTTAAAATCTTACCTTTTCTAATTTAAAAGATGA
CCTCCTATTTTCTAAAATTACCTACATTCCTTTTTGAGAACGGGTCAGCTGTGTTGTCC
TGTCTGGCCTTGAACTTGTGAACTCAGATGGCCCTTCTGTATTGTCCTCATCAGGAGCT
GGCTCTATTTTCTAACATTATTTTTACTTTGAATTTATCCAAGTCTTCTGCTATGTCAA
GATTTTTCCCAATGCTGGGGATCAAATCCAGATCTTCGAGCATGCTAGGCAAGTGTCCC
ATCACTAAGCTACCTTTGCCGGTTAAGTCAAGTTTTGATGCAACCTAAATGGCTTAAGC
ACCCTGAAAATAAAGATTTTGCCACTTACAATGTGTCCTTCTATAGTCAATGTTAAAAT
AGCTTCTAAAAACCCAGAAGCAAACCTGAAGACACAGAATAACAAGTAGTTTTCCACCG
GTCTTTGGTGAGTCTGACTATGACATGTTTAAATAGCATAGGAAGTTAATTTTTTTCTT
CCTTGTAGTTGGTGTCTAAATAGTAATAATGTAGGCTGTGATTGAAATACCTAGTTCAG
GGAAATCATCATGTTGCTGGCCATGTTTGCAGGGCTCCCAACAGAACCCTGGCCGTGTA
CAGCATGTACAGTATAACTGAGTATGTTACTTGTGTTTTGTCCAGAGGTACTTGACAAA
ATTTTATGACCTTAAAGAGGACAGAATTCAAAAAACAAAATGGAAAGCCAAGAGAAACC
TCCTTGAAGAAAAAATCCAGGAAATGCAGCAGTTCTTTGGGCTCAAAGCAACTGGGCAA
CTGGACAGCCAGACTCTGATGATAATGCACACACCTCGATGTGGAGTTCCTGATGTGGA
GAATCTCAGAGAATTGCCGGGGATGCAAAAATGGACGAAGCATCACCTCACCTACAGGT
AATGTTCGTATACACCCTAGTGAATTCTACTTTCCCATGGGTGTGGCTTCTAGAAACCT
GAAACATTTAGCCACAACAAAGTTTTAAATCTAAGAGTAGCCTCTGTATGCTGTTAAGG
TGATGGCTCAATGGAGAAAGTGCTTGTCATACGTGGGAGTCAGAGTTCAGATCCCAAGA
GTCCATGGGAACACCATTACAGGCATGGTGGCTGTCTGTAATCCCAGCACATGGAAGGC
AGAGATAAGAGATGCCTAGGGCAAGCTGGCTAGGGAGACTAGCAGAATCAAGGTCTCTG
GTATGGTTGAGAGATTCTGCCTCAATAAATAAAGTGGAGCATGATTGAGAAGACACCTG
AGTCTACTTCAGGGCTCTACATGCATGCATGCGCACACAACACACATACCTACCCAGGC
ACATCCCCCTACACACACAATTCATTATCTCTATAAGGACAGTTGTTCACTGAAGGAAG
ACTTGCAGTCATTCTCTTGTCAGTCAGTCCAGTGCAGAACAGAAAAACTGCAAGCAAAC
CAAACAAAACTGAATCGAATACAAATAAAAGACAGAAGTCTGCAGAATATTAAAGATGG
ACAAATGTGGAGTTTGGGCACAATGCCTATTTGTTTTTATGAAAGCAAATCCTCTTCTA
TGAGGTTTCAACTGAAGAATCGATAAAGAAAATGTGGTACATTTATACAACAGAGTACT
ACTCAGCTGTAAGAAACAATGATATACTAAAATTTACAGGCAAATGGACAGAACTAGAA
AACAAACCATCCTAAGTGAGGTAACCCAAAACCTCACTCATAAGTGGACACCAGAAATA
AAGCAAAAGATACCCAGCATACAATCTACAACCCCAGAGAAGCTTGAACCCTCTTCCAG
AAACAGATGGAAGAAGATGAAGAAATCCACAACTAACCATTGAGCCAAGCTCAATTGAA
GTGAGGGAGGAGCAAAAATATGAACAAAGAAGTCAAGACGATGATGGGGAAACCCACAT
AATCACCTGACCTGAGCTAGTGAGAGATCACTGACTCAGGGAACCTCCATAAGACCAAA
CTAGACTCTCTAAATATAGGTAACAATGGTGTGACTGGGACGATGTATGAGGTCAATGG
CAGTGGGTGCAAGATCTAACACTAATGCACAAATTGACTTAGTGGAACCCATTCTATAT
GGAGAGAACCTTGCCCAGCCTAGACACAAGGGTGTGGGGATGGAGAGGTGCCTTGCTCC
TGCCTCAACTAGATGATGGGACAGACTTCACAAACTTCCTAGGGAAGACCTTACCCTCT
CCATGGAGCAGATGGGAGAAGGTGGGTTGTAGGGGGAGCAGGAGGAGAGGAGTGAGGGG
GAACTGCGATTGGAATGTAAAAAATAAATTAATAAAAAAGAAAGTGAATCCTTCACAAT
AATTACTAGTAACTTATAAATAAGAAGAAATTAAATTACATTTTGATTGATAAAGGTTT
TCCAAAGAAAAAAATGGTTTGAAGTTCAACATTTTTTTAACATGGACTGTCCGTACGTG
GGGAAGTTCTTCTGGCAGGCGTGGGTGCAGAACTAAGAGACTAGAAGAGATGAAATTAA
CCAATAATGTCATGTTTGTTTCCCTCTTTCTCTAACACAGGATCTATAATTACACCCCT

GACATGAAGCGTGAGGACGTTGACAGGGCATTTCAGAAAGCTTTTCGAGTCTGGAGTGA
TGTGACCCCGTTGAGATTCAGGAAGATTTATACAGGCCAGGCGGACATTATGATACTGT
TTGCATCTGGAGGTACACAGCTTCACCTGTGTCATTTGTCAGGCTGTGTGACGGCATCA
TTCAACACACCCATGATGCCACCAAGTGTTGTGCTCTGTGATGGCATACCTTCTGACAG
ATGTAGCTGACTAATGACTTTCTTTGTTTTAAAAAGCTCATGGAGACTTCAGTGCTTTT
GATGGCAGAGGTGGCACAATAGCACATGCTTTTTACCCTGGACCCGGTATCCAAGGAGA
CGCACATTTTGATGAGGCAGAGACCTGGAGTAAAGGTTCTCGAGGTAGGAGAGTTTTCT
TTCTTTTTTTTAATGAAAACTAACCATAACTAAATTTAACATAGGCCCTGATCTTGAAA
GAATACATTTTGATGAAATATTATAGTCTTAAATGTTAAATTTTGATTGTTTACCTCTA
AATCAAATTTAAGATTTTATATATTTAGTCCAATGTTTCCTCATTGCAAACTGGGGTAT
TGGGTGTGATTACGTTAGTGATCCTATGAGCAAGTTCTGTACAGTAAACTCTTGGAAAG
GGTTTTGCCAGTAAAATCTATAAACTTAAAAGCAAACGCTAATTCTGTGATTGATTGTA
TAGTGACAATTTCTTAGATTCATGATGACGTGGCAAGATATTTATAGAAAATTTACAGA
TCCCTATGGATAAGCTCTTAGGTAGTCTGGGAGCCTTGCTTGATCTTGGGATACCATTT
TCAGAATGTCCAATCAATCAATTGGAACAGGAACTCACTATACCACTCAGATAATGGGG
AATGAGGTTAACTTATTATTAATTTATTATTATTATTAGAGGCCATCCCTAGGAAAAGA
AAATTTAAGCCATCATTTCAGTTGCTTTTCTAATATCTCTTGGAAGTTCTCAATTATAC
AGTTCATTCATGTGGAAAACATAGGGAGACATTCAAGAATGTGGTCTCTCTACCCTGAA
TCTCCCTTCTCTGACTGATTTTTCATATGATGTATAGCTATATTTTTACAAATAGAAAG
GTAGACACATAGACTACTGCTTTGACACAGCTTTGATAAAGGCATTTTGAAGGCTGAAA
GGTAGATCATGTATGATTTGGCACTTGCTGAACTACACGTTCTCAACCAAATATGGTCC
ATGTGTGAGATTTTCTCTAGGGTCATTTGGGGGCTGGTTGCAAGTTCTTCACCAAGTGA
CAATGCTTTGGAATTCTAATGGACCGTAGAAACGTAAATCAGTGCAGTTGCTCTCACCT
GAGGAGAACAAGCAAAAGTCAGTAGCAAGTTTGTGTATGACTAGAACATGTACTATTGG
CTCCTGACCTTCTCCCCCTGCCAAGACAGGGTTTCTCTGTGTAGCTTTGTAGACCAGGC
TGGACTTGAACTCACAGAGATTCTCCTGACTCTGCCTTCCAAGTGCTGGGATTAAAGGT
ATGTGCCACCACCACCTGGCTATTCCTGGTCTTAATTTAAAAATTGATTTGACTTTATA
TTTTAAAAAGGGTGTTTTTGAAAATTCTATCTGTGTGGCCCAGGCTGACCTTGAACTCA
TGTTCCCCCTACACCATCTTTCTGAGTTGGGATATGGGCATATAGTACCATACCCAGTA
TTGATGTATTGTTTTGAACCAGTAATGAAGATTCTGATAGGTGCTTGATGAGTTTCTGT
TACAGTGTTGAGTGACATTTTTGATTCAGAGACAAAAAGGACAAATCCTCCCCAGGAGC
ATGCAACTTAACAGAAAACCAGACAAACAGAAAGTCCCTGTTTGCAGATGAACATTGAT
GGGAGTTAATTTTACAAAACCTTATGAACTTGAATTTCTAGACTTATCCAATCCCCCTG
TTTCAGTCTCCTGAGTGGCTTGGACAACAGACACATGCAATTGTCCCAGCTTTAAGAAA
GAATTTGAAGCAATGTGAGCTATCAAACATACCTGTGATTTATGTTAGGAAGAATCATG
CCAGAATGTTTCACTTATCTGGCAATGAGATGGTCCATAACTATAGACAATTTTGGCTC
CAAAGGGACATTTTCAGTTAAAAACAGAAGTGGGGATTCATACTTCCACAGAGTGGTGC
AACTAGAGATCTCTAAATATCTGTAATAGAGGATGGCTCTGGCAGTACTTACCTAGCCT
ATAATGACAACACTATTGAGGCTGAGAAAAGCTATTTGAAAGAGCTGGGAAACAAGCAA
AGGAGCCTCAAACATCCCCTCTACTGTGTTAGATGTTGTATTATTTCCTTCCAAATGGC
AAGGCAATGGGGTGGAAAGATGGCTCAGAGTGTACCTAGCTGGTAGAGGGCCCTCTTTC
TAGTTCTATTCCCAGTACCCAGGTTGGACAGCTCACAACTGCCCAACTGTGGCCTCTGC
AGGCACCTGCACTCACATGGCACAGCATAGTCACAGGGGGATGACATTCAGCCACTAAA
GGATTGGCTCCCATCCCTTGAATGGAACTGGGTTTTCTGTAAAAAGGTCAGGCATAAGA
AGATGAGCACCCTATTATCTCATTCATATTGAGACATGAGAAAAGTTAGTCTCACAAAT
GTTACAACAGAATGGTGGTTCCCAGTGGCTGAGGAGAAAGGGGCGATGGATGGGTGATG

GCTGAGCCCTTGGTGCTAAGTCACAGTTAGGAGCATGGAGCCATAGTCTTGACACTGTG
GTGCATCAATACAACTATAGACAATAGTAGTGTACTATTGATTTCAAAAGGCTGGAAAG
AAAGGATTTTGAGTGCTTTCACCACAAAAAAGTGGTACATATTTGAAAAGATATATTTA
TCTTGTCTTGAACATTATGCAATGAGTGTAGCCATCAAAATATCACATAATATCCTGTA
AACATGTACAATTTCATGTGTGAGTTAAATAAAAAGTGATAATCCATTTCACAGATTTA
TTCTTTTACCTCCTTGCTTCTCAGGCACAAACCTGTTCCTCGTTGCTGTTCATGAACTT
GGCCATTCCTTGGGGCTGCGGCATTCCAACAATCCAAACTCAGTAATGTACCCTAGCTA
TAGTTATGTTAACCCCAACACATTCCGCCTTCCTGCTGATGACATACAAAGTATTCAGT
CTCTCTATGGTAAGCTGAACCTCATAATTTGCCATACTATACCCTGTGTTCTTATGAAT
AATCTCATTTGTATATTTAGAAGATTTTAGTAATTTAAATGTCATGTGATAGGCTGCAA
ACACTATCGTAATTTCGCCTCATATTTGGAATGTGTTCTAACAATGGCATCACTGATCT
GTTGATGGGGCCTCCGGAGAAAAGAAATTCAATTATTCTCATCTGTGATGCTGCCACTA
AGTGGGACAAAGTCCTCAGGGCAGGCTTCTTATGAATTTCATTGGAAGCATGACTGTTT
TGTATTTTCTTTTAAGAAACAGAATATTTAAAAAGAATAGCCTTTCCTCTTACAAAATT
TCCCCAAATGAACATCAAAATGTATTTTCAGGATCCCCAGTGAAAAACCCACCCTTGAG
AAATCCTGTCATTCCATCAGCAGCAACTATCTGTCAGCAAAGCTTGAGCTTTGATGCCG
TCACAACAGTGGGAGACAAAATCCTTTTCTTTAAAGACAGGTATGGTCATTTTCATAGC
TAATGAATGCTACAGGATATGTCGCTAAATCTGCTAGTTTTTTATTTGTTTGTTTTTTG
TCAGAAGATATTTGAAAGGATGTAGATCAAATTTGTAAGGTAAATTTGTAGAACTTGAT
TCCCAAGTTTTAGAGAGATCTTAGATTTTTTTTTTTTAAATAGTGAATTGGTGTGATGC
CATTTGTGTTGACTAGGAAGCTAGTGTGTGAAAAAAATTGCTTTGTAGTAGGAATCTGG
CTGGGGAATTAAAAGACAGGAAGTGGAGAAGCCCAGGTAAAAATGATGAGGTCTCAGAG
GTTGAGGAGATGGTTCGGTTGGCAAAGTATGCAGTGTTCAAAGAAAGCCTGATATTAGT
CCCCAGCCCCACACAGAAATGCCCAGGGGCAGCTGCATGGACTAGGAATCCCAGTGCAG
GGAAGGAGGGCAGGGTTGCTGGGGCTTGTGGAGTCTAACCCAGTCTAACTGAACCCAAA
AATCCAGGCTCAATGAGAACATATGTCTCAAAAGATAGAGGGTGTGGAGTGATTGAGGA
AGACACTTGTGGTCTCCACATACATGTTCACAAACACATATACATATACATGAAGACCT
TCCCCCCAACACACACACACACACTCACGAACTCTTGAAGAAGATCATGTGTAGGCTTC
CTTGTACGTGGAGAATTCCAGAATGACTCTAGTTTCTGTTTTGGAGCAGAGATGTAAAC
AGATGCTGGCTAGCAGGGAGAGCAGTGGGGAGGGCTAATGCGTGGGACAGAGGATGGAT
TCCTGCTTTTGAGATTATAAAACAGTAATTTCTTCAAAACCAGATAACCTTACAACCTT
GTGAAGCCACACCAGAATCTTAGATCCCTGGAGTCCAGACCCTGGGTGTTTTTTTGGGG
GGGGTTCCTATTTAAGTCAGGTTTTCTTTCAAGGTGCTCGCACACAGAAATCCCTGGAA
TAGAATGTGCAAGAAAAAAATACTGAATCATTTCACAAGTTCTCATTTGTGTTTTTGAA
GAATACAAATCTACAGCAGTTTTTCACATTTATGAAATGCGTTTCCCACCAGGTTCGTC
TGGTGGATGTTGCCTGGGAGTCCAGCCACCGTCACTTCAATTTCTTCCATGTGGCCAAC
CATCCCATCTGGTATTCAAGCTGCTTATGAAATCAAAGGCAGAAATCAACTCTTTCTTT
TTAAAGGTAACTCCAATGCTTGGTTCTGCCACTGAGAACACTTTAAGGAGGAGAGAAGT
AAAGGGAAGGGTATCAATTCTGTTATCAGAGATCTGATTGATAAGTTCTGGGATAGGTG
TATAAAAGATGAGTCAATGGCACTTAACACATCCAAACCATAGAACCAGGGAGGTAAAC
CAATGATTGAGAACAAAAAAGACTGGGGTGACAATTTCTCAGTGACAGAGAACCTATTT
ACTATGGATGGAGTCTCAGGTTTGCCTCCCACCCCCAAGCACAAATAAAATAAAATTAA
TTTTAGAATTCTGCACAAGAGAGATGTTTTATAGTAAGTCAGTATTTGTATTATTCTTT
GGTGTGATTTTGTTGTGTTCTTTGTCTTTGTCCTTATAGATGACAAGTACTGGTTAATA
AACAACTTAGTAGCACAGCCACACTATCCCAGAAACATATCCTCTCTGGGCTTCCCTGC
GTTTGTGAAAAACATTGATGCAGCCATCTTTGACCCATCTCATCATAAGGTCTACTTCT

TCTCGGATAGACGATATTGGAGGTGAGGTCCAATGGCAGGCACTTTGTCCACAAAGATT
TGAGAAGCAGGCACATCTTGGAATTTTCAAGAAGTCAGCTTTCGTTTGTTTGTTTGCTT
TTTAATGATCATCTGCATATTTAAAATTTTTATTTTATTGGAGTTTTTGCCTACAGGCA
TGACTGTGTGAGGGCACTGGATACAGAAGGCAGCTTTTCTATGAAGGTTTCTAATTTCT
TTGCCTAGGGATAGGCAGCAAGAGAACACAGGTCTGAAACAGTCACTATCGACCTGGTT
CTTTACTCTGGGCCTTTGGTTTTTATTTAATTTCATTTTGGAGAGTGTTTGGACTCCAG
GACTCCTTCTCATAACCAGTGGACCCAGAAGACAACATGAGTAGCCTGCCCTTTGTGTG
ACAAGAGAGCCACAAGGAAACCAGGGCCTGGTTATCTTACTCACCTAATGTTGAAAGAG
GGAGGGATCATCCCTACGTTCTTCTGAAGAAAACAGTTTGCATACTTGGCATAAGACAT
TAACAGGTTATACTAAGATAATAAAGCCTCTATGTTTTCTTGAATACTTAAAATTCAAT
ATCTAATTTCAACTGAATGCCATATATTTGTGACATGAATTGGTTTGACAACTTTAGCC
TTGTTTGAGAAATAGGACTTGTGATTTCACCAAATAAGACTTTCTGAATAGGTGACAAA
CAGGTGCAGTTTATTGTAATCCAACATTCTTTATTTTGTTGTTATTGATGATTGATTGA
TTTTTCAAGACAGGGTTTCTCTGTGAAACAGTCTTGGCTGTCCTGGAACTCACTCTGCA
GACCAGGCTAGCCTCAAACTTACAGAGATCTGCCTGCCTCTGCCTCCTGAGGGCTGGGA
TTAAAGCTGTGTATCACCACCACCCAGAGTAATCTAACATTCTTAGAAATTTGTCTGAC
TACTGGTTCCCCCTTTATACTTCCAACTGACAGGGGAATTCCTTTTCCTAAATTCCAGA
GGGACAGTGTCAAGCAAACAATGGTTGGAGAACTAAACTGAATGAAAAATAAGACAACT
ATGTTTCAACTGTCATGTGGAAGTCTAAAGAAAAATTGACTCTTTTCACTCAGTGGTAA
AAGAGACTTTATGGAAATACTAGGGCCCCCCCACTCTCCAGTGTGGACATGACTTGTTT
TACTTTTAGTTTTTCTTTAACATGGATGTTCGAAACCCATTGGTTCCATGACTCTTAAG
TACCCCTTCTCTTCACTGTGGGAAAAAATCATTGATATGTGTCTTGCTAGAAAAGATTT
GCAAAAAATGGTGATAGACAACTAAGTCATTGGGCCACTGAGGAATGTCGCATATCAAG
CCCTCTTCAATGCCAGGCTGTGTGGGAAGGGACTAAGCAGTGGGGAAAGCACTGTGAGA
ATGGAAAGAGATTTGACTAGAGTCAAGGTTCAAGTTTTGACCCTTGTAGTTACTTAGGG
AAAACCAACCAACTGTTTATGTCTTTGCATTTTTCATCAGTACATTTAAAACTCTCTCA
GGAAAGACCAACGTGGGCTTCATGTTATGAAGCCCATTTTCTCATTGTGTTCCTGCTGT
TTCCTCCTACAGATACGATGTGAGGAAGCAGTCCATGGACCCTACTTATCCCAAACTTA
TTTCCTTGCACTTCTCAGGAATTAAGCCTAAAATTGATGCAGGCTTCTCTTTCAAAGGT
AAATAGAGTGGGAAAATATTGGAAATCACATTTTAAAGTGTGATACCACAGACTGGGAT
CCAGGGTGGGTATAAGATGCTTGTGTTTGTGGTGTGAATATGTTCCACTTCAGCCTGAG
CAAACTTGCCTATCAGTGTCACATAGCAACAATACAACCCACATTGTGTGAGATGAGCT
GCATGACACGTTTAACAAGAATGGGAACCTCATTTTATTCCTTCTCCTCTCCTGTTGAG
TGGACCAAGATTCAAAAACAAATTTGAGTTTTCTCTAGCCTTTTCCTCCTTCCAGATCA
CTGTATGCCTGACCTAGATCCTAGTCTTGAGACCTCCACCCAGAGAGTGACTTGCTCAG
GGGGGAACATCAGAGCAGGGTGCACTGCCAAGGGTCCCCGATATTGACATCTCAGTCAC
ACCTTTTGTAGTCTAGATAACCTATAAAATTAACTCTATGTTTTAAAGGAAGACACGAT
AAGAATTAAAGAATTGAATTTGTTTCCATTTTAGGACACTACTACTTCTTCCAAGGAGC
CAATCAATTGGAATATGACCCCCTGTCGAATCGTGTCACCAAAAGGCTCAAAAGTACGA
GCTGGTTTGTTTGTTTGAAATGATGCAGTTGAGGATCTTCGCTAGTTCTTCAGTTTAAT
AAGTATTTATCACATCTGCACTTTATGCTCATTATGCATATAATGTAACATGAGATAAG
GTGAAGTGTACAGGCCACAGATAAATATTTACACCGAAAAATGCTTTGACAAAATTTAT
CCTCTTCTGGTAAGCTTTTTCACTTGACTCCTTTCTTACTTTTGAAAGCGGGTACCTGC
CAAGTCTGCCAGGTTTCTTTCTAAGTTGTTTTCTAAGAACCTTCAAGTGCACCAATACG
AATTACTTCCCTGTCTTTACTAATATTTAATGTGTATTATTTTGTCAAATAAAATGTAA
AGAATTTAGCTTTTGATTTTTTTCTTTAAAAATTGTGTGTGTGTGTGTGTGTNNNNNNN

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTGTGTGTGTGTGTGTGTGTGTGT
GTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTGACTGCATGTAA
GTATGTACACCACATGTGTGCAATGCTGGTAGAGGCCAGAAGAGAACATCAGATCCCTA
GAAACTAAAGTTACAGATGGTTATGGGCAGACTTGTGGGCCTTGGGGATTGAACCTGGA
TCCTCAGGAAAAGCAGCTTGTGCTCTTAATGTCTGTGTCATCTCTCTGGTCCTGTTTTT
TGTTTGTTTGTCTGTTTTTGTTTTTCTTACTATTTTTCTTGAGGAACAGAATGAATTTT
GAAAAAACTCTGGCTTCTAAATAAAGTCTTTCTGTGACTGTATCAAAGAGTCAATAGAG
ACGTGATAGCTACAAAAATATTGAAGCTTGAAGTCCTTTTGATCAGTTGTATAAGGACA
CCTGTTTAATCCTATGACAAGAAAGCATCCTCTAGCCTGCACATGACCCCAAGAATAAC
TGCAGCATGTGACCACAAGACTTGAGAGGAGAGAGGATGAAATGGGATTTCCATTCTTG
ATAGATATGTTGTGTGGCTCATTGCACATCAGAGAGTGGGCTGTGCCATTGCTGTGGAG
CCTTACAGGGTATTAAATCTATGACTCCTTAAAAATAGGAGGAAATATGAGTAGCACAC
ACCTCTAAGTTTGTTGTGGAAGAAAAACACAACATGGTGATCCCCTCACACCTGTGTCA
TGGCCTTCTAGCAGTTTCCCGCATTGCCCTGACATTTAGGCCCCTATCCTCTGCATTGT
GTGTGCTGTAGTGCTCTCATTTAAAGCTATGGATTTGGGAAATGTGTCCAATATTAAAA
CTGACTGATGTATTGACCCAGATACAGAAAGCCCTCCATAGCAGATGAACATCTGGTCA
TTTTTGAAGAGTGAAATGAACTGAAGGTGGAGAGATGTGGGGAAGGGGCCCTCCAGCTT
TGATGTCTCTGACATTGTGAGTGATGATATTCACCATCAACTTGATGGGATCTGGGGTC
ATATAACAGACAGACACCTGAGGACTATCCAAAGAAGGTGGACTGTGGGAGGAATTCCC
TCTTTCAGAGTGGTGGCGGGTTGACAGCATTAGAACACATCCATCATGGAAAGGGCAGC
ACTTTGTCCTTACAAGGATAGATGCTTGAATTCTGGTTATGAGTTTGTACTTCATGTAC
AGAAGCTTCTGCTCAAACTGTTGCCCAGGGACTTGCAGAACACCTTATCCACCGCCATG
CCTTACTGCACTGCATTGTTTCTGACCAGGGAACTCATTGCACAGCCAAAGCAGAGTAG
CAGTAGGCTCATGTTCATGTTCTTGGAATTCACTAGTCAGATCATATTCCACACAGTCC
CGACACATTCAGCTTGATGGAACACTTGCTGTCTAGGCTGCTGTAAGCCTTTGATTATT
TTCTAGATTTTTTTCTTTTTGGTGTTAGAGATGAACCCAGGGCCATGTGTATGTTAAA
AAAGCACTGTCCCTTTAAGCTATACCCTGCCTGTTAGGATTCTAAAAATGTTGACTTTG
CCATTAACAGCTAGTTGTTGCTTTTCCGAAAAAGGTATTTTGGTTGACCTTATTCTATT
ACCCTAGGTGACATGGCTCTGAATTAATTTTAGTGTTGACAGCAAAAAGAGATTTAAAG
CCCTAAATTGGTGCTCACAGTAACTGTCAAAATAGTTATTTTGAAAGCCATATTATTTA
TTTTTTATTTCAATTGTATTTATTAATACCATATAGTGAT

9.3 SERPINF1 GENE SEQUENCE IN CHINESE HAMSTER GENOME

55000..80000, scaffold 942, reverse complement

GGAACTAGCTCTTGTAGACCAAGTTGGTCTCGAACTCACTGAGATCCGCCTGCCTCTGC
CTCCCAAGTATTGGGATTAAAGGATTGTGCCACCACGGACTGGCTTCCCTAACTTAAAA
AAAATTAATTTTTAATTATATGTAGGAGTGTGTGGATATGTGTCCATGAGTGCAGTTGC
CCTCAGGGGACAGAGGAGGGTGTGGAACACCATGGAGATGGATTTACAGGCAGTTGTGA
AGCACTTGATGGGGTGCTGAGAACAATACTCCGGTCCTCTGTAAGAGCAGTACATTAAC
TACTGAGCCATTTCTCCAGCCCCCTATCGTGTGTATGTTTTAATTTTAACATTTATTTA
TTTTGTTCATTTGATTTTTGAGTTTTGGTGTTTTTTGTTTGTTTTGTTTTGTATTTGAG
ACTAGGTCTCACTATATGTAGCCTTAGCTGGCCTGGAACTCATTATGTAGACCAGGCTG
TCCTCAAACTCACAGAGATCTGCTGACCTCAACCTCCCAAGTGCTGGGATTAAAGATGG
GAGCCACCACGCCTAGCAGTTTACTGACTGACTTACTTATGACTCTGTGTGCCACAGTG

CACAAGTAGAGCTGACAGTCAACTTGCTGGTTTTAGCTCTTTCCTTCTACCATATGGGT
TTCAGGGATCAAACCTGGGCCCTCAGGATGGCAGCAGACTCCTTCACCTGCTTAGCCAT
CTTGCTGGCCCTTTCCCCTGGCTTTTAATTTTCTATAAATGAATCATAAAGTCTTCTCC
TCTGTTGGGTGCCCTGCCCTGGCCTGGCAGTGCTGGGGATTGAACCCAAGGTCTTACCT
ATGAGGCAAACACCCCACCATTGAGCTGTTTCTACCTCAATTTGTATTCTCTGTCGGGT
TTCTTGCACTCAGCCCCGTGTGTTGGAGACTTGCCTGTGTTGTGCCCAGCAATGCCCCT
TCACCTGGAGGCCTTGGGCTTGGAGGTTACTGCTGGATCTGAGTTCTGTGGCTTCTGGA
TGGAAAGGGATGAGGTGGGAATAAGAAGCTGGTTGGCTGGCCCACCCAGGGTGCAGATC
AAAGCCATAGAATTCTTCTTGGGAGGGAGGAGAAGGAAGTACTCAGGGCAATGGTATTT
GGCATTGCCATTCCTAGGGTTTCTTTGTAAAGGAACGGGTAGGGGTGGTCCCTTTGGCT
GATCTGGGAAGCAGAAAGACTATGGAGTGGATGGGGGATTTGCAATTTAAACTCGTTAT
AATGGAAATGTTATGCTAATCCAGGCCAATCACACGTCCCTTTTTGACTGGAAAAATTA
GGACTATTCCTGTCATTTGAACTAATTAGTAGCATTCAGTCTTGCTGTTGGAAGTGTCC
TATAAATTCATTCCTGGGCTACCATGTGGATCTGGTAGGCATGGAATTGCTAATGATGG
TTTTGAAGGTAGCCATTTTAAATGTCAAGGGATTGACTGAATGCTTTGGTAAGGGGTGG
GAGTGTATACTTGACACATAGAGATGCTTTAAGTCATGGAACTAGGGGCTGCATGGGGC
TCCTGCTTAACTCCTGCAAGGGCCTGGGTTTCATCCCTAGTGGTGTGGGGAAGACACTG
AACAGGAAGAAATCCCTTAAGGGGAAGACAATGTAACAGGAATGAAAAGGGTCCCAGAT
CAAGCAAGCAGAGACAGAGATGGGAGGACATGGGAACTGTAGAGAGAATGGGCTTCAAG
AAGGAGGAACACTGCTTTGCTGCCTGCCGCTGAGATTGGAGTTTCCATGTGGTAGGGAA
CACCAGTCACCCAGGTGATGGCGGCAGTGTTAGAGAAGGTGGGCGCAGAGGCACAGAGG
GGAGGAGGAAGCGGGAGCAGAAACTATGTAGACACGAGAGCAAACAGCTCTTTCCATCA
TGGCTGTGAGGGAAATGGGCAGTGACCTTAGGGGATGTAGCATAAAAAGGAGACTTTCA
GATAATAAAAGTTTTTCTTTCTTTCTTTTTTTTTTCTTTCTTTTTCTCAATGTTCAGAT
TGAACCCAGGGTCTTGTATGTGCTGGGGAAGTTTGCTCTACACCAGGCTATACCCCAGT
CCTAGGAGATGTCTTTAAATTTTTTTTATTTGCACATTCTTGTAGGTTTATGTTTGCAT
GTGAGTACATGTGCCTTAACACACAAGTAGAGGTCAGAGGACATCTTCTGGCTGTGGAC
TCTTTCTGGCTTGGCAGCAAGTGCTAACAGCTGTGCCCTGGGAGTGCAGACCTGTTTTG
AAGCCTTGAGTAACTTCTTGTAGCAGACACTAAAAGAAGAGGCTGAGAGTCTGTTGGGT
GAACTTTGGATGTTGTCTAATTTCCGGGAATCCCACAGAACCAAGAATGAAACAATGAG
CTTTCTCCACTCTTCCAGTAGAGGGCGCACTCTTGCCAGAGCCTTTGGCAGCCAGTTGC
ATCCAGCATGCTTTCCCATCATGCATCTGTGATGGTTGGCTTGGTGGGAGCTACGATGT
AGAGAGCCCTCTTTACCAATCCCCTTCTTCAGGTGCAGCCTGGGGAGGCACTGGCAGTG
GAGTAGAGCAACTGGATTGAGGTCACACTAGAAGCTTTCTTTAAAAAATCAAGCAAACG
CAACACTGGAAACAATGCAGCTATACTCCCTGCGAATAATTCCTTGTTTGCATATGCTC
GTTTTATTCTTTGGTGGAATCGGGGATGGGATCTAGGACCTCATGCATGCCAGGCAAGC
ACCCTTTCAGAGCTTTATCTCCAGTCCTGTGGGAGGCTTTTGGTGCAGGACTTGTAGAA
GTAGAATATTGAGGCTGGGCAGTGGTGACTCACTCCTTTAATCTCAGCGCTTGGGAGGC
AGAGGCAGGCGGATCTCTGTGAGTTCGAAGCCAGCCAGCCTGGTCTACAGAGTGAGTTC
CAGGACAGCCAGTGCTACACAAAGAAACCCTGTCTCGAAAACAGACAAACAAACAAACA
AAAAGAATAATATTGGAAGTCAGCTTCGGGCTTGTCCATTTTTTTTTGTTTTGCTTTTT
GTTTTTTTTGTTTTTTGAGACAGGGTTTCTCTGTAGCTTTGGAGGCTGTCCTGGAACTA
GCTCTTGTAGTCCAGGCTGGTCTCGAAGTCACAGAGATCCATCTTCCTCTGCCTCCCGA
GCCCTGGGATTAAAGGCGTGCGCCACCAGCACCCGACGTCTCTTACTTTTNTTTTTTTT
TTTTATTTAAAACTTGTTGTTAGATTTATTTTATGAGTGTTTTGCCTGAGTGTGTATCT
ATGCACCACATGCGATCCTGGTGCTCACAGAAGTCAGAAGAGGGCATTGGATCCCCTAG

AACTGGGGTTATGGAGTTGTGAGCCACCCCGTGGGTGCTTGGAAATTTGCTAAGCCAAA
TCCTCTGCAAGACTAAAAAATGCTCTTAACTAATGAGCCATCTCTCAATGCCTTTTCCT
TTTTGTTTTTGGTTTTCTCATTACTTTACAATGTTTTGGTGTGGTTGTGTTTTCTCTTT
GGTTTTTTGTTTGCTTTTTTTTTTTTTTTTTGGACAGAGTTTCTCTGTGTAACAGCTCT
GGCTGTCCTGGAATTAGCTTTGTAGACCAGGCTGGCCTCAGACTCACAGAGATCTGTCT
GCCTCTGCTGGGATTAAAGGCATGTGTTACCACTGCCAGGCGTTTCTTTGTTTTATGAT
ATCTTTGTTTGAAAAGAGAAATTGCTAACTTTAATGCCACATTCATCATCAACCTTTTA
CCTTACAGTTTGCACTTCTTTGGGCATTATTTATTTTATTTTAATTTTTAGAGATGGAG
TCTCACTGTGTATTCCTGGTTGGCACGGAATTCTTTATGCAGGTCCCTAATCTTAAGTT
TCCAGCACTTTTCCTTCTGTCTCCTGAATCCTGGAATCACATGTGTGCTACCACACCCA
GCTATTTATTTATTATTGTTTGTTTTACCTACTTGGATGTATATGTGTACGCCATGGAG
AATGCATGTGGAGATCAGAAAATCACCCGAGGAAATCATTTCTCTCCTTCCACCAAGTG
AGTCCTGGGGAATCCACCGGTTGTCATCAATCATGCTTAGTGTCAAGTACTTTTGTTTG
CTCAGCCATCTCATAGGCCCCATATTTTTCTTTCTGAGATAGGGTCTTATGTATCCCAG
GTTGGCCTCCAACTCTTTTTTTTTTAAGGTTTATTTATTTAATTTATGTATGAGGACTC
TATCTTCATGTATGCCTTTATGTCAGAAGAGAGCATCAGACTCTATTATAGAGCCACCA
TGTGGGTGCTGGGAATTGAACTCAGAACTCTGGAAAAGCAGCCAGTGCTCTTAACCAGT
GAGTCATCTTTCCACCACCGGCCTCCAACTCTTACATGGAGGATTGCCTTAAAGTCTTG
ATTGCCCTTCCTCTACTCTTCAAGAACTAGGATTACAGGTGTTTACCATCCTGTCCTGC
CTGGATTTGCTTTTTAGAGCATGGTATTGCTTAAAGAATATGCTGGGCGTGATAGGGCA
CACCTTAAATCACAACACTGGGAAGGCAGAAGCAAGAGGATCACTGTGAGTTCCAGACC
AGCCAGGGCTACATAGTGAGGTCTGGTCCCAAACCAACCAACCAGGGAAGATATCAAAG
TCAGCTGTGCTCAGAATGTGACACCCCTGGGTGATATGGGAGATTAACTGAGAAATTTT
CACGGGATAGTTTTTGACTCTCTCTCTCTCTCTCTCTCCCCCTCCCCCCCCAGTGCGTG
CGTGCGTGCGTGCGTGCGTGTGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG
TGTGTGTGTAGGTCAGAGGGCGACTTTGTGGAGTAGGCACTCTTCTTCCACCTAGGATT
GAGCTCACATCACCAGGCTTAGACTGCAAGTGCCTTTACCCGCTAACACTAGCACTTTG
CCTAGAGTCACACAGCAGTTTTTCTTTATTGAAAGGACCTGGACTCTGACTGTTGACAC
TCTCCTTCCTATGAAGTTGAACCTTGCCCACTGCAAAATCTCACTGGTCCCAGCTCTGC
TTTCTGGGCTTCCTCGAGAATGCCAGCAGAGTCTCCTACATTTTCTCTCTTTTGAGCCT
TTTCCTCTTTTCTTCTCTGAATGGCTATCAGGCTTTCTTAGTCAGACTAAAAAACATTT
GCTATATAGGTCTAGAGAGCAACATTTTAAACCAGAGCAGCTGTGGAATACGAAGCATT
CCACAACCTTCAGCTCTCTGGAGCAGATCCCATGCTACACCGTTGACAGTTCAATGTGA
TGAACTTTGTCCATGAGCAAGAGTAACAAAAATATTGTTCTTTAAAAAAAATCATTACA
TTTATTTATTTTGTGTGTGTGTGTATGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG
TGTGTGAGAGAGAGAGAGAGAGAGAGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAG
CCATGGCACACACAAGAAAGTCAGAGAAACACATGTCAAAGTTGGTTCTCTCCTTCCTT
ATGGGTACTGTGGATAGAACTCAGGTTTTCCTAGTGACTACTGAGACACCTCACTGGCC
ACTTTTTCTTTTCTTTATATTTTATGACACAGGGTTTCATGTAGCCCAGGCTAGCCTCA
AATCTACTGTGTAGCTAAGGCTGGCCTTGAACTTCTGCCACCACTTTCTAAGTGTTGGG
ATTATAGACACCATACCAGTTACACAAATACTACCTTCAGGCTGTGTGTATTAAGAGTT
ATATGAAGGAGGTGAATTTCAAATTTAGAATTTCAAGTTTCATTCCCAAGATCTTTCAT
GTAGATGCAAATACTCCAGATCCCCAATTCAAAACTGTTTACTTTGGATCCTAGGTGTT
TCAGATAAGGGGATTCAGCTTGGAGTCTCTAGTTGTCACTTGTTGACCTTTAGTTTTGA

AATTCTGTTACAAGTTCTAATACCCCAGGGCAAACTGAATGCAAGTGTTCCCCTGTGCA
GCAAACCTGTAGCGTGGGCCTCCTCTAGTGACAGGTGTCCTGTACACCTCATTTATTCC
TCTGCTCAGATGTTTCCCCCACTGTCCCTTCTATTCTTCTTAAAGGTTTTAGTGTGTGT
GTGTGTGTGTGTGTGTGTCTGTATGTGTGTGGTACCTGAGGAAGCCAGAAGAGGGTATC
AGATTCTCTGGAGCTGGAGTTACAGGCAGTTGTGAGTTCCTGGTGTGGGTGCTGGACAC
AGAACTAAGGTCTTTTCTAAGAGCAGCCAGTGCTCTTAACCACGGAGCATCTCTCCAGC
CCCCACCATCCTCTGTAATCTTCAGAGAACACGGAACTCCAGAACAGACAGTCCCAGTC
CATCCTTTTGCTTTCCAGGACAGGGAGAATTATCTTTCTCCTCTACTGCAGGCAGAGGC
AAAGCTAACCACTTGGAAGAACCCAATGTCTCCATGCAGCCCACCTCTCAGCCAATCCT
TTCTGATGTTCAGGTTCAGATTCATGACCTACCCCAGGCTGCCTCAAGTGCCCTGTGAC
GTTAGCCATCATGTGGAGGGTCCCACTGAGGAGGGATGAAGGCCCAAGGAAGATGTGGT
GGGAGAGGAGGGTGACATGAATGTGTGGTCTTGGAGACATAGGATGGATGGGCAGCAGT
GGGTTAAAGTTAACACCAGACAATGATGCAATCACAGACTCCAAATTGAGTCAGGTCAC
TTTAAGAAAGGAGTAGCTGTAATCTGAAGCCTGCTGGACGCTGGGTTGGGAGGCAGTTA
TTCACTCCCCTGCTTGCTAGAGCCCCCTCAGGGTGCAGGCTGAGAGGGACCTAAACTCA
GAGAGGAGCTGCTGTGGACAACAGGTAAGGCAGTTCCTGGTCTAGGCTGGAGAAGACAG
ACGGGACAGGCCCCTTGGCCAGAGGGACAGGGAAGAGCAGGGGCACCCCAGAGAGCAGA
AAAGAGGGGTACAAGGAGGTAGAGGTAGGGATATAGTAGCCTCTGTACTTTAGGGACAA
AGGTAAACAGATGAGAGGAAAGAAGAGAATTGGGGGGCGGAGGGCCGGGATACTGTTAT
CCCCTACATTATCCCACTGCAACAGAAGGAGAGGCTGCATGGGGGTTTAAATGGGGTGG
AAAACAATGTGGGGGTATGGTGGTGGGCAGTGGGGAGGGCAGCTCTGGTCTGGTCTGAA
CAGTATCCCAAAGGGGTTATCATCTCACAGGCAGGGCTCTCCCTGGAGTTGCTTCCACC
AGCAGTGAGCAAGGAAGGTGGTCTTCCTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCTGCCTGAGTCCTG
CTGGGATGGGATGCCTGCTAGGTCCTCTCTCCAGCTCCCACCCATTCTGTTACCAAGTG
AATGGGAGACAAATGAGGTCATGGCCTCTTGCAAACAATAAGCACCATTTAGCCTAGCT
GGTGTGTGGGGTCTGGCTGCCCTTCTGCAGTCTTTTTACCTCTCGAGGAAGCCACAGCA
GGGAGTGGTCTGGGTCAAAGAAAAGAGACAAGCTTTCCAAGTGCCAACCACTATCTGGA
CTTTAGCTGCCCCCATCAGTTCAGCCTCCAGGCTGAGGGGGATGGGTGGGGACCAAGGA
GTTCTGGCTCCAGGGAGTTCTGGCTTTTCTGGGGTGCTGGGAGGTGCAACTATGGGGTG
TTTCAGGTTCCAAGATTTCAGAAAAAACCAAACCAAACCAAACCAACAAAACTCTCACA
AAACTTCCCCAGAGGTGAGTGGTGTAGCCTTGAGCCAGCAGAGAGGGAGGTGTCCTCCC
CTGCTCTGTCCACATGTCCACTGCTGTGTGGCCTGAAGGCGGCTTGGTTCTGGGAGGAG
GAAGCAGTGGAGGAGGAGCAACATGTTTGTCTCCCAGCTGGGTTTAATCTGAAACACAT
GTACTGGCTCCCGTTTGTCTTCCCAGCTTGGATTTGAGACCACAACATCCCAGTTTCTG
TGCCTCTTGATTTGGGGCAGACCTTGGCTCAGAGGGGACAATCTTGACCCTTTCCCATG
CCCTCCCCAACCTGTAGCCCAGGAATCTGGCTTGGACGCTGCTCTTTGGACAGAATGGC
ATCACTGTCCCAGTTCCCCTGGCTAGTCAGTAGCTGCCTACCCAACCCAGTTATAGCTC
CCAGGGCACTGTATGTGTCCTGGGGCCCCTCTTGCAGACCCTTACTTCCTGCCTTTCAG
GGTCTCCACTTCCTTGATTCCTGTGGCTTCAGGTGAAGATTTCACACTTCCATTAACAC
TTGCCCTGCTTATAGCTCTTGGGCTAGCACTGGACTCCCATGTCTCCCCATGCCCACCA
GGTTCACTACCTCAGCTTCAGTTCTCAGCCTACCCATGTGGTGGTGACAGTGATCTGTG
ACTTAGACCCTGGAGGAACTCTGGCAGGCCCCTTCTGCAAGTGTATCTCACTTACTTCC
TGCAATAGGAGCCAAGGGTGACCGTTTTCTCTCAGTCGATACCCACCCAGAGGGACCAG
AGGGGGCTTTCTGTGCATCTTTGATAAATCAGCCTGCTGAGGCTCAGCAGGGGGAGGGT

TCTGCCTGTGATGGGTGAGGCTGGGAGGGGAGGGCTATAATCTGTATCACTTTTCTGTT
GTTCCAATGGTTTTTTTCCCCCACGAAGCTGGGAGGCAGTGCCCTGGGTGAGAACACAG
GGGTCCCTAGAATTCCCACTGGGAGATCTGCCATAGATGAGGGGTTAGAACAAGGGAGG
CTGGTTTCTGCTGGTGCTGCTGCTGGAAGCAGAGGAAACTTTGTCTTGTTTCCCCTCAT
GGGCTACAGTTGTCTAGTGGATTATCTGACTTAATTCTAAGAACTTAGCCAATGACTGC
AGATTGCACTCCAAATTTTTCAGATCCAATTTGTTCTGGGAAGTCAGAAGAGTTGATGG
TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGGTGGGGGAGGGCAAGTGTATGTGAC
AGGATAATGATGGAGCTGAGGGAGCATTTCTGGGTATGTGTGCTGTAGGATGGGCGGAG
GACACGCCATGCCACGACAGACAGTGGTCTTTTCCGTCTCTAGAATAAACTTTTCTGTG
GACCTCGTACCCCAACCCTAGACCTGACTCCAGGAGCACAGGAATTTTGCTTTGTGTTC
CAAGTGCCAGACCAGAGTCTGGGCAGAAAGGTCGCTGCTTCGGGAGGTGATAGGAGGTC
CCACCAAAACTTAGTGTTGTTTAACCTGAAGGAAGGAAGGAAGGAGGGTAGGAAGGAAG
GAAGTTAGGAAGGAAGGCAGGCTGACAGGCAGGCTTTGGGGTTAAGAGCTGTTTCCTTA
AAGTGGGGGAGAAGACATAGCCCTTCAGAAGAGATGCCAGTGGAAGGGACAGCTGAGAT
CTTCCCAACCCCTGGCACTGCTAAGGGCTGCTAAGATACTATCTAGGCAGATGTGGTTG
TGCACATTTAGGAGGCAGAGACAGGAGGACCACACATTCAAGGCCAGCTTACATATAGC
AAGACTGCTTCAAAAAAGAAAAGAGAGCGAGAGGTTATATAATATAACGCTCATAACAG
TAAGTAAATACTACTATATATGGAGGGGTTTGTTCTTTGGACAGCTATGAAGTTGGGGT
TTGATCCCAGGCAGTGTGGGCCTAGTTTGCACCCTCTTAACTGAGTAAGCAACTGCTAT
TGAGATGGGACACTGTGGTCAAGCCAGGCACAGTGGGTGTCAAACAGTGGGATTCCAGC
ACTCAGCAGGCAGATGGAGGAGGGGAGGAGGGTAGCTGCAAGTTCAAGGCTAGCCTGGG
CTACACAGGCAAGATCCTATCTAAAAAANNNNNNNNNNNNNNNNNNNNAAAAGGAG
AGAGGGGCGAATAGACTGAGTGTCCTTATCGTGAGCTGCTTTTGGCCAGAAAACTTTTG
GTGGATAGCAGGGCTAGCCCAAGGTTCTGTGAGTCTCTGAGCCCCACCAAATGAGCTGG
CTGTATGCCTCCATGAAAAAACCTGGGTGTGTGTTTATTTTCTTGGGGCCAGAAGAGTC
CACTGCTGGGTCTTCACTGCTCTGGCCCATTCTGGCCCTTAGGCTAGAAGGTGGGTCTT
CCTATAGGAACTATTGTGGAGGCACAGCTCACGCTAAAGAGGAGGCAGGTGTGGGTGGG
GACAGCAAACCTTAAGCTTCGAGTCCTTTTGGGGGCTGAGAGCTGGAGGGAAAGCTCTC
ATCCTGTCTGACTCTTGTCCCTATGAAGGGTGGGGTGGCAGCACAGGCGGTGGGAAGGG
GAGTCACTTGCCCTCTTGTTTCTCCCAGGATGCAGGCCCTGGTGCTACTCCTCTGGACA
GGAGCCCTGCTTGGGCATGGCAGCAGCCAGAATGTCGCCAGCAGCTCTGAGGTCAGTAG
GGTGGGCAGCAGGAGTTCTAGCTCCTCTTCGGAGCCAGCAAAGGGAGGCAGCGCTGGGT
CCGGGGTGGGCTCTGGCTTTCATTCCTTTAGTTGGTTTCTCCAGTGGCAAGACAAATAT
TTGCCTTACCATTCAGGCTATGGAGAAGCCTCTCCTGTGCAGAGGAAGTAGGGATAGGG
TCTTGTCACACAAGGCACAAGCCAGTTGGGTATATTTAGTCTGTTTCCTAACAAGGCCA
CTATGGCCATTCACTTATTCCTTATCTAGTAACCACCTGCCTCAAAACGCAGGACCTTT
CACATGCTACACAAGTGCTGTATTAGCCCTGCCCATGAACACTTATTCTGTGGTTGGTA
GGTATGGATAGAAATATAAAAATCCCAGCATTCAGGTGGGCAGTGGTGGCACTCGTCTT
TAATCCCAGCACTCAGGAGGAAGAAGCAGGTGGATCTCTGTGAGTTTGAGGCCAGCCTG
GTCTACAGAGTGAGTTCCAGGACAGCTAGGGATACACAGAGAAACCCTGTCTTGAAAAA
CCTAAACAAACAAACAAAAATAAATGAGGCTAATTTTGCTTTTAATTAAAATATGTGTT
TGTTATTCTTTTTTATGGCTCTTGTCATTTTTTTTTATGATTTTATTTAGTTATTATGT
ATACAACATTCTGCCTCCATTCTGCACACCAGAAGAGGGCACCAGATCTCATAACGGAT
GGTTGTGAGTCACCATGTGGTTGCTGGGAATTGAACTCAGGACCTCTGGAAGAACAGTC
CGTGCTCTTAACCTCTGAGCCATTTCTCCAGCCCCGTGTTTGTTATTCTTATTATTATT
ATCATTATTTTGTTTTTGTTTTTCAAGACAGGGTTTCTCTGTGTAACATTCCTGGCTGT

CCTGGAACTCACTCTGTAGACCAGGCTAGTCTTGAGATCTGAGATCTGCCTGCCTCTGC
CTCCTGAGTGCTGAGATTAAAGGTGTGCACCATGACCACCAGCATATTTATTATTCTTG
TGTGTGGGCACCTGTGGCACAGCATGTATGTGGAGACTAAGAGCAACTCTGCAGTCAAT
TCTTTCCTTCAGCCTTCACATGGGCTCTGGGAATCAAACTCAGGTCCCCAGGCTTGAGT
GGCAAGCTCCTTTGCCTGATGAGTTCTATCACCAGCTCCTCAGATTTTACTTTTAAAGA
ACTCAATTCCTGAGCTTGCTGTGATGGTGCACGTCTTTAATCCCAGCACTCGGGAGGCA
GAGGCAGTGGTTTGATGATGCCAGCCTGATTTACAGAGTGAGTTCCAGGACAGCCAGAG
TCACGTCTTACCACCCTCCTATACCCCCAAACAGCATCAGGAATGGGAGATGTAGCTAA
GTGTCAAATCACTTGTCCAACATATGCAAGAGCCTGGGCTTGATCCCTAGCCCTAGAAA
CAAAACAAAAACTAGCAGAGTCCCTCACATAGCAGCTAGCTAATAAGGAAGTGGTTGAC
ACTGGAAAAGGAACCTAGGCAGACTCAGAACAGTCTGAAAGGGTGAGCAGGAGTGTGGT
ATGGCCTGAGGACAGGTTTCTTTGGTGGTGGTGGTGGTAGTGAGGTCTCCTGCCGAGAG
ACAGATTGTGTTTGTAGGCCAGAAGTCAATCCACAGGACCCCTCACTGGTCTGGAATAC
ACTGACTTCTCCAGGACCTGCCCATTTCTTCCTCCCCAGTGCTGGGATTACAAGCATGC
CCAGCCTTTTTTGTGTGCCTTCTGGGGATTGAATCCAGGTCCTCATGCTGACAGGCTGG
CACTTTCCCAACTGCGCTAGTTCCCCAGCCCTGCCATTTTTATTTTAAAGTAGGCCAGG
CTCCCTGCTCTCAGGTCACAGTTCAAGCTCGGTGGACCCTCTTGGTTGTCTCTTCCCAG
CACAGAGGTTATGTACAGTCAGAGTGCACATGCTCACTCCAACCTTAGACTAATGGACC
CCAGGCACCAGAGGACAACCCCAACCTGGAGGTGTTGCCATTCACTGTGTCTTCCCCAC
TGGTCCTTGTATTCTGACCCAGGGTTCCCCAGCCCCTGACAGCACAGGGGAGCCAGTGG
AGGAGGAGGAGGACCCCTTCTTCAAGGTCCCGGTAAACAAGCTGGCAGCAGCAGTCTCC
AACTTCGGCTATGACCTGTACCGCCTGAGATCCAGTGCCAGCCCAACTGCCAACGTTCT
GCTGTCTCCACTCAGTGTGGCCACAGCCCTCTCTGCTCTTTCTCTGGGTGAGTGTCAGC
CGAAGGAGGCTGCAAGTGAACCTGAATTTTCCCACGGAGTCTCTATGCATATGCTGCCG
GGAGAGCAGGAAAGGGACAGTGGGGGCTACACCTCTGGCCAGGCACTGGCCACACACAC
ACAATCCCTACCACTTTATGAAAGTGACACTCAATTCTTGGTTCCTAAGTGGACTCCTA
TTGTGTGTTAGTCATTTATCCAGGTTCTAGGGGTACTGCCACGAGCAGGCCCAGGTTCC
TCCTCAGAATAGTGTAGGTAGCAAAGGGAAAACTTCAGGGACTGCACCCCACTCTTCTC
TCTCTGTAGGTCTGATGGCCTTGGCCTATTTTAGAAATTTTTCCTAGCAATGCAGTTAT
ATTTTCCTTGAAGCTCTCATTCACTTATATAGAATGTCTGCAATGTCAGCCTCCTTGGG
CTCTCAGAAGACAACAGACTAGGTGTACACTATCCCTGTCCTGCTCTGGGTCCCATGGA
TAGCCATGAAGCCTGCCATCCTTTTCTCCCTGGCAGGAGCTGAACAACGAACAGAGTCC
ATCATTCACCGGGCTCTTTACTACGACTTGATCAGCAACTCGGACATCCATAGCACCTA
CAAGGAACTCCTTGCCTCTGTTACTGCCCCAGAGAAGAGCCTCAAGAGTGCTTCCAGAA
TTGTGTTTGAGAGAAGTCAGTAGCCCACCCACCCCACTCCTGAGTCTGTGTAGTCCAAG
CTAGTCCCTGACTCATAGTGCTGCTCCTGCTTTCAAGTGCAGGCATGAGCCATCATGCC
TGACACATGGTTGCTTTTCTGGTCTGGGTTGTCTGTGTGCTTCATGCTGAGGGTTCTGT
CTAAACAATGGACTGTGTGTGCTCAGGCCAGGCCTGGTGCTGGTGGGCAGGAAAGGGCC
TGTGAGAGCACAGAAGATCGGAAAGAGGAAGTGGACTGTGGGAGACAGCAGGAGGGTCC
AGGAGGGCTGTGACATGAAGGGGAAGGCAATGGGTGGAATCCATTGTTCCTTGTCCAGG
GTTTGGGCAAGTAGGCAGATGCCCCAGCAAGTCAGGGGAAGATGCTGAGCAGGAAGCTG
GGGTCTGGAATGCTTTCAAAGTAAGGGTGCTTGTTGAGCTCTCACACACTGATCTTTGA
TCTGCTTCCTCTCAAGAACTTCGAGTAAGATCCAGCTTTGTTGCACCTCTGGAGAAATC
ATATGGGACCAGGCCCAGAATCCTCACTGGCAACCCTCGGATAGACCTCCAGGAAATTA
ACAACTGGATACAGGCCCAGATGAAAGGGAAACTTGCTCGGTCTACAAGGGAAATGCCC
AGTGCCATCAGCATCCTCCTCCTCGGTGTGGCTTACTTCAAGGGTGAGGGCTTCCCCAC

TTCTCTTGGATGGCAGGTGTGTGGTGGTGGTGGTGAACACAATGCAGGTGCAAGGCCAG
AGGAGGGTATGGAGCCATGTGTGCTTGTTCATCTCATCATCTACCAAGTGTTCCTGCTC
CATGTGACACGTGAATGGCTGGGTGGGACTGGGCACAGATCTCTGACCAGGTTTAATCC
TGGCTGTGCCCTGTCTCCTCAGCTTTCCTGTGAGGACTACATATCCACTTCACTGTACA
ATACGGGCTGTCTCATTCTCACAGAGACCCTGTCCTGCTCCAAGGATTCCACCTGTGCT
CAGGACAATGCTTCCAGTTTCCTGCCAACACTCTTCTAGGCAGTAGTAACGTGTCAGGA
GACCCTTCCACAGGCCTCATACAAGGTCTGTTTTCTTCAGAAAAGGGCTTGCAGCAGGA
ATGCTTTGACCAACTTCCCTCTGAAGATGTGACTGCCTTACTGAGTAGCAGCAAGAGTG
GACCACCTAACCGAGTAGGAAGCCCCTGGAGCTTCCTACTCGGTACACAGGGGATAGGA
GAGCAAGCACAACCCAGCATATTAAAAGCCACCTTAAACAAAAGGCCAAAGTGAGAACA
CCCACTCACTTCCAACTGAGTTTGCCAGGGGAAGAGTGTCAGCACCTGTGGACCACGGA
TCCTGTGGTCTACACCACACATGCTGGGACCCAGAAGCTCCTGTGTCCATGCTACTGAC
AGCTCAGGTCCCCTGGCCTTTCCGGTGGCATTGGCTCAGGAAGCTCCATATGTTTCCTT
CCAGGGCAGTGGGTGACAAAGTTTGACTCGAGAAAGACGAGCCTCCAGGACTTCCACTT
GGATGAGGACAGGACTGTGAAAGTCCCCATGATGTCAGAACCCAAGGCCATCCTACGAT
ATGGCTTGGACTCTGATCTCAACTGCAAGGTGTGGGAGCATGGGGGGTGGGAGGGGTCA
GAGAGGGGCAGGGTGGTATGGAATGGATGGCTGCTGAGATGGGTCAGCTATCTGAATTT
TGCTGTTGTGACTTTGGGCAAGTTAGTAATGTTTCTGGGCCTTTCCTTTTCCTCCTCCT
CCTCCTTTGTTTTTTTTTTTTTTTTTAAGATTTATTTATTTATTATGTATACAGTTATCT
GCATGCATGCTTGTACACTAGAAGAGGGTACTAGATCTCATTACAGATGGTTGTGAGCC
ACCATGTGGTTGCTGGGACCTCTGGAAGAGCAGTCAGTGCTCTTAACCTCTGAGCCATC
TCTCCAGCCCCCCTCCTTTGGTTTTTAGAGACAGGTTTCTCTATTTAGCTTGGGAGCCT
GTCCCAGAACTTGCTCTGTAGACCAGGTTAGCCTGGAACTCAGGGATCCACCTGCCTCT
GCCTCCCAAGTGCTGGGATTAAAGGTGTATGCCACCACTGCCTGGTATCTTTCCTTTTC
TTGAATCTAAGGAGTGAAAAGTAGTTATCAGAACTTATTCTCTAGGATTTGTTGTAAGG
ATCAGAGCCATAAAACACAGGTTTTTAAGACCTGGCATTCAGAGTTGGGGATGGAGACA
CACACCTGTAGTCAGTCTCAGGAGTCAGGAGCCTGAGGTAGGGTGACCAAGAGCCCAGT
CTGAGCTGTACTTTGAGACCATCTCAACAAACAGAGATGGGCTGGAACAATGACTCAGT
GGTTAAGAGCACTAGCTGCTTTTTCAGGGGAGACAGGTCTGATTCTCCAGCACCCAAAT
GGCAGCTCACACCTGTTTGTAACTCCAGTTCCAAGGGATCCAACACCCTCTTCTGGACT
CTGTAGACACCAGGCATGTCTGTGGTGCACAGACAGACATGCAGGCTATCACCCATACA
TATAAAAACAAACAAACAAACAAACAAACAAACAAATGCTGGGGTGTACCACAATGGTG
GAGCACGTGCCCAGTACGAGCAAGTTTAATCCTTAGTACTGCCAAAAACAAAACAAAGA
AAAAGACTTGACACTCAAATGTTGATTTATTTATTCATTTCTTTCCTTTTTTATTTTTG
AGACAGAGCCTCAAACTTACTAGCTGAGGATGACCTTGAACTCCTGATTTTCTTGCCTC
TACCTCTCAAATGCGGCCTTGTAGGCATGAGCTACCACCCTGGTTAAAGCAGTGCTGGG
GATCAACCCACAGCTTCATGCATGTTAGGCAAGAGCTCTACCAAGTGCACACAGCCCCA
GCCCAAAGCTGTCTTTCCTGACAGCTCTTCTTTTAGAATTTAGACTCAAATTCTGACAG
CTTGTCACAAACACCACACAATGACATTCTAAGGCTCTCAGTGAGCTTGACTTCTTTGT
TGTTGTTGTGTTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNTCTTGTAGACCAGGCTGGTCTCGAACTCACAGAGATCCCCTTGCCT
CTGCCTCCCGAGTGCTGGGATTAAAGGCGTGCGCCACCACTGCCCGGCTCTACTGTTCC
TTTCTAGTCCAGGGATCTGAACAAAGCAAATGGTCAGTAAATGCTGCCAGGGAGATTCT
GGGAGGATATCCAGTTTCAGCTCCTAGGAACCAGGTCCCGGACAGTGTTACCAATTTGA
CACAAGTTAGAGTCTTTGGAACAGCAGCTCAGAACCAGGCTGGCCTCTGGGCACACCTG
TGGGGCATTTTCTTGACTAAGGATTGATGTGGGAGGGCCTAGCCCACTACAGACTATGC
CATTTCTGAGCATGTGGTTCATGGCTGGTCTGAAAGGGCAGGCTAAGGAGCCCGGGGAA
GTAAGCCAGAAGCAGCGTTCCACTTTGGTCTCTGTTTCAACTCCTGCCCTGAGTTCCCC
GTGGTGATGAGCTTGACCTGAGAGTTCTAAGATAAAATAAACCCTTTCCTCCCCAAGTT
GCTTTTGGTTACATTTTATCACAGCAATGGAAACCCCAAGTAAGATAAACTATAAGCTA
ACCATTTTCCATCAAACTTCTGAGAAGAAAAGTGCTTACTCCTGGAAGGACTTCCAGAG
ACCCGCAGAAGGCAACTGCTGCACTGGCTGGGGGTGGGGCCTGCCTCAGGGGAGGGGAA
AGCAAGTCTTAAGGTCTGGGGTTTTGGCTCCAGACTAACTCACGTGCTTCTCTGTTGTT
AGATTGCTCAGCTGCCCCTGACAGGAAAACTGAGCATCATCTTCTTCCTGCCCTTGACG
GTAACCCAGAACTTGACCATGATAGAGGAGAGCCTCACTTCTGAGTTTGTTCATGACAT
AGATCGAGAGCTGAAGACCATCCAGGCTGTGCTGTCTGTCCCCAAGCTGAAGCTGAGCT
ATGAAGGGGAAGTTACCAAGTCTCTGCAGGAGATGAGTATGTTCTCAAGACCACTCTAC
TCCAGGGGCTGGACATAAGGCTGAGTCTGTGAGCCTGCCTTCTGCCTTGCTGACAGAAC
TTTTGGGTGGCTCTGGGCACTTGCTACACTGTTATGGGGGAGAGGAATATAGAACCACC
TTCCCATCAGACTCAGTCCTAGGTCCNNNNNNNGCACACGTCCTGTAACAGCCCAGAAT
ATTGTCTCAGGCCAGGTCTGTTTCCCTTGCTGGCTCTATGCTCTCAGGTAGGTCACTGA
GCCTGCAGTGGCTTAGAGAACCTCCTACCTCACAGGGAAAGTTTGCTTGCTTTCCTGAG
AATCAGCTTACACAAACCAGGATTGTGTGTGTGTGTGTGTGTGTGGTGTGTGTGTGTGT
GGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTCTCTCTCTC
TCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTGTGTGTGTGTGTGTGTGTGT
TTGCTTAGGAGTATGCATGCATCTGTGTAATTGAGTATGGGTGCCCATGTAGGCCAGAT
GTGTTGGGTCCCATGAAGCTGGGAGTTACAGGTGGCTGTGAGCCACTCAGTGTGGGTGC
TGAGAACTGGACTGAGGTCCTCTTCAAAAGCAGTATGTGCTCTTAACTGCTGAAACACC
TCTTCAGCCCTTTTATTTTATTATTTTTGTTTTGTTTTGTTTTCTGAAACAGGGTTTCT
CTGTGCAGCCCCGCTGTCCTGGAATGTGCTTTGTAGAACAGGCTGGCCTAGAACTCAGA
GATCTGCCTGCCTCAGCCTCTTAAGTGCTGAGACTAAAGTTGTGCACCACCATGCCTGG
CTTAGCTGTTTTATTTTTGAGACAAGATCTTTTGTAGCCCAGACTGCCCTCTAACTCTT
GACCATCTTGTCTTGGCCTCCTGAGGGTTGAGATTATAAGTGTGCTGCTACCATGGCCA
ACCTATGCTAGATTTTAATGCAAGCACAAACATTCTCAAGCTAAGGTCACATGAGACAA
TTAGAGATTTCTTCTTTGATTTTTATTTCATGTGTGTGCGTGTTTTGCCTGCATTTATA
TCTGCGCACCATGTGCACACAGTGCCCACAGAGACCAGAAGAGGGCAATGGGCTCCCAG
AACTGGAATTACAGAGAGATGTTGGCTGCCATTTGGGTACTGGAAATCAAACTCAGGAC
CTTCAGAAGAGCAATACATACTCTTAGCCACTTAGCCATCTCTCCAGCCCTGAGAGACA
TACTTCATAACTTCAGGAAATAATTCTTTATTTTGGACCGGGATAACAAAAGAGTTGTC
TGCAGGACTAAACTTTTTTCCTGAGGAAACTCTATCAAAAAAACCTTACAGGTTGGACT
GGAGAAATGGTTCAGAGGTTAAGATGACTGGCTGCTCTTGCAGAGGTCCTGAATTCAAT
GCTAGCGACCATCTATAATGAGATCTGGTGCCCTCTTCTGGCATGCAGGCACACATGCA
AGCATAATATGCTATACATAATAAAACAAAACAAAACAAAACAAAACAAAACAAAACAA
AACAAAACAAAAAAACCCAAACACCTTACATGTCACCTCTTTGTAATACTACATTAATA
GACAGCAGGAGAATCAAATTCGGCCCACTTTCCATTCCACAAATGTTGGAATGCTATCT
TCCATCACCCAGCTTCTTGGAAAGATGCTGTCAGGCATCTTCAGGACTATTTATTTCTT
GATTGTAATGCCATGGCAACTAACAGCCACTTCTTGGGCACAGCCAGGGTGGGTAAGTT
CCTAAAAGCAGTGTATAACTCATCTGACCAAATTGGTCACCAGTAATAATGTAATAAGG
ATCAGTCAGAAAATCCCACAGTTTCTGACCTTGGGGGTCTGGACCACCACTACACATCC
TTGAGGAAAGGGGAAGCCATCCCTCCCTTCTGTCTCACTTTGAAACCATAAGCAGCAAT

ATCTACAACCTCAGCACTTGGGGGTGGGGGGCATTTGGGTATCAACCATGAGCCTCACA
CTCTACCAGTGAGCTACACCCCCAGCTCCTTGTGGAATGTTTGTCGGACTCTTGATTCT
CTTTGTACACAGAACTTTTAAACCTAGTACTTGGAAGGCTGAGGCAGGAGGATGATATG
TTTAAGGATAGACTGGCTTACAGAATGAGGTCCTGTCTTCAAATAAAAAAAACCTGAGT
AGTTTATTTTCAGGTTTGCTAACATATAACTCAGAAAGGAGGCATATGTGCTATAAAGT
CATCCATCCCTTCAAGTTTGAGTTTTATTAGTGCCTGGACAATGAATCTGTCAAAAAGA
TTTAGTGAAAGCCAGAAAATAAAATAAAAAACTGGGGGCTAGAGAGATGGCTTAGTGGT
TAACAGCACTGGTTGCTCTTCCGGAGGACCCAAGTTCAATTCCTAGCACTCACATGGCA
GCTCACACCCGTTTGTAACTCCAGTTTCAGTGCTTCTGACACTCTCACACAGACATACA
TGTAGGCAAAACACCAATGAAATTAAAAACAAAAAGAATTAGAAAACCAGCAGGAAACA
AATACAGAGTTGCCATCCCAGGAAGAAACTACCCAAGAGTAATATAGCCCCATGTTTAT
TTCTCTACTCACAAATTATACACCACCAATTTTATTGAAGAATTTAAGGAAGCAGAATA
TTGTAAGTATAATTGTGTGTCTCTTGAAGTAGAAAGTCAACCAATTTCTTGCTGTATAC
TGTTTTTTAATTACAATACAAGACTAAGAACTGAAGTTCCTCCTGGAGTGGTACACAGC
ACTTCATTGCTACCAGAAACCATAGGGATGTCCCAGGGAAACAGACTCTCACATTTAGT
GAGATGAAATCAGACTCTCACATTTAGTGAGATGAAATCAAGTTGAAACCCTAGGCTAG
TCTCAGCTGCCTGCAGACCAGCGGCGACCTTTAAAGCACTGGGTTTCCACAAATACTGG
TCTAAGTCTTTTCTGGGGATGGGGAGTTTCTATCAGAGGGCAGTTAGTTTCTGCAAGCA
ACAGCCATCAGGACAGTGCCTCTCAACCATTGTTTCCCAAGTGCTCCCTGCCATCAAGA
CACAGCTTAGGACAGCTGGGTAGGTCCTGTTCTTTCTGATGACAACTTCAAACCAACAT
AGCTTCTTTACCTCATCTGCCAGAGTTCCCTTCTTGTGAGCTGGAATGTACATTCTACA
CCAAAGGCCCCACAGGGATGAATGATGAGCCTAATAAACAGGATTTCATCTCCCGGAGC
TCCTGGACCTCATCGCTCTCAGGGCCACAGTGAACCAACAGGATCTTTTCTGCTTTCCA
TATGGCATTCAGAGCTTCAGGCACTGCCAACCTGAACCAAGGGAGGAAAGCTGGTCAGT
TCACTCATTCAGTCAAAATGAAAGCTGGCTCCTTTCTGGACAGAGGACACAAAGGATCA
AGTTTAATTCCATCCTCCACAATCTACAGAGGGCACAAAGGGTTCACCATACCCTCCAG
AATCTGATCTTACCATCTAAAGGGGTAGATAAGAAACTTCACAAGCACACTGCAAAGCA
GAGTTTGAGGAAAGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGAGAGA
GAGAGAGAGAGAGAGAGAGAGAGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG
AGAGAGAGAGTGTGTGTGTGTTAGACTCCAGGATCCTGAGAAAAGTGGGCAAATTTATG
AAAGCTTTATACAGAGTTGTTCATTTTAGGGCATCCATGTCCCACCTGACTGACTTTGT
GTTTCCATGTCTGGTAATTAGGTCTAAGCTTCAAACACTTTCCCTTAGCAATGCACACT
AATGAGTTTTTAATACAGGCTGGAGGTATGGTATCCACACAGGAACACTACAATTGGGG
ACTCTTTCATTTCAAGTGATCACAGTAGCAAAATTGGGTGTAGGACCAGCTAGTTCATC
AACTCTGCTCTTGTCTTTTTTTTTTTTTCCCTTCCTCTTCTTCTGTGCAGTGCCAGGGA
CGGAACCCAGGGCCTTGTGCATGCTAGGCAAGTGGGCTACAACTGAACCATTCCCCCAG
CTTCCTTTTTTAACCTTTAAAGCAGCAATTCTTAACCTTGACGTAACTACTTGGTAACT
CTACTGTTGAGGGTGTCCTGTGCACTCAGGGTTAGCCTTTACCCACTAGGATGCTGGCA
GTACCCTCCCTCGGTCCATAACAACCAAAACTGTCTCCAGACATTGGCCAATGCTATCT
GAAGGCAGAATTTCTTCTGCTTGAAAACAACTGCTTTAAAAGGAATAGGACGCAAACTT
GGGAAAAGGCAGAGGCCCTGGAGTTGAGAATGCTGGGAATGAGTAAGCCAAGGGGGTGT
GATTAAGCCCAGAGGAATGCTAGAGGGCGGGTGAGGCTGCCTGGAAGCCCTCCCTGGGA
AACTCAATTGAGAACTAAGAGCTTTCTTCACATCTCTCAGGAGAAAATTCAAAGCTTAA
AGCCTGTCAGTGAAACAGAGGGCAACGATCTGTCCTAAGGAGGAGACAGAAGGAAAGGT
TTCAACATCTGAAAATCACTGTAGCAAAGTCATCCACAAAAGCA