GENOMIC AND TRANSCRIPTOMIC APPROACHES FOR THE ADVANCEMENT OF CHO CELL BIOPROCESSING

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF
THE UNIVERSITY OF MINNESOTA
BY

Nandita Vishwanathan

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Adviser: Professor Wei-Shou Hu

June 2014

Acknowledgements

I wish to express my deepest gratitude to Prof. Wei-Shou Hu for been a major contributor not only to my thesis, but also to my life. His energetic and enthusiastic demeanor has always been a source of inspiration to me. I also really appreciate his patience and mentorship during the harder times of my PhD. I am grateful for all the opportunities that I was given, especially for the opportunity to coordinate the Consortium for CHO Systems Biotechnology.

I would like to thank Prof. Nathan Springer, Prof. Benjamin Hackel and Prof. Robert Tranquillo for taking the time to serve on my thesis committee. I would also like to thank all my teachers at the University of Minnesota for being passionate and giving. I had a great time through all my courses, where I also learnt a lot.

The Hu group has been my home away from home. I will sorely miss the Thanksgiving and Christmas parties at Wei-Shou's. Sheau-Ping has always been a gracious host and has a very warm and caring heart.

My teachers and colleagues Andrew Yongky, Dr. Yonsil Park, Dr. Kartik Subramaniam, Dr. Shikha Sharma, Dr. Anne Kantardjieff, Dr. Kathryn Johnson, Dr. Huong Le, Dr. Nitya Jacob, Dr. Bhanu Mulukutla, Dr. Siguang Sui and Dr. Marlene Castro-Melchor have always been there for me. Huong has been not only a great friend, but also a mentor and a fine role model. Her discipline has been a source of both my envy and inspiration. I have to thank Nitya for looking out for me every time and Bhanu for having great discussions and debates and for all that I know about metabolism. Siguang has always been open hearted and ready to help. My sincere thanks to Kathryn for always being keen to help, and also for lending an empathetic ear to my frustrations. I will sorely miss my days in the lab, and can hardly imagine my life otherwise.

I want to thank my undergraduate students Yutin Chen and Baizhen Gao for their assistance, and especially Xiaolu Zheng for her hard work, brilliance and diligence.

I want to thank my colleagues Mohit Sharma, Dr. Hsu-Yuan Fu, Ravali Raju, Sugandha Rajput, Arpan Bandyopadhyay and Tung Le for being loving and for allowing me to boss over them, especially during the short courses and consortium meetings. I will

always be indebted to you for your kindness. Dr. Mingyong Xiong, Dr. Liang Zhang, Dr. Hsu-Yuan Fu and Dr. Kyoung Ho Lee are brilliant scientists I have had the good fortune to associate with in the past few years. Merely being in their presence has taught me a lot.

I wish to extend my sincere gratitude to Jessica Raines-Jones, Kaitlyn Pladson, Kirsten Keefe, Erin Fenton, Jenna Novotny, Lindsay Bork, Lamisa Chowdhary and Shruti Saxena for lending a helping a hand during the busy days. The department staff, especially Mary Nissen, Julie Prince and Teresa Bredahl are very loving and eager to help.

My thesis is a result of many rewarding collaborations that I am grateful for. I have to thank Prof. Daniel Voytas for welcoming me into his lab where I had the great fortune to learn and interact with Dr. Feng Zhang, Dr. Yiping Qi, Dr. Kim Nguyen, Dr. Colby Starker, Dr. Yong Zhang and Dr. Nicholas Baltes. From this lab, I learnt the basics of molecular biology and 'some cool tricks' that I showed off to my labmates. I have to thank Prof. Scott McIvor, Prof. Nikunj Somia and Prof. Nathan Springer for allowing me to drop by and for enthusiastically sharing their perspective on our sometimes-crazy ideas. I have been fortunate to work with great industrial collaborators Dr. Yung-Shyeng Tsao and Dr. Zhong Liu at Merck. I would like to thank Dr. Getiria Onsongo, Dr. Kevin Silverstein at the Minnesota Supercomputing institute; and Joann Mudge and Thiru Ramaraj at the National Center for Genome Resources for their guidance. My special thanks to Faraaz at the Bioprocessing Technology Institute (BTI) in Singapore for being a very efficient collaborator. You sudden demise has been a great loss to all us. I also would like to acknowledge the efforts of Terk Shuen Lee, Sze-Wai Ng and Dr. Bernard Loo at BTI.

I would like to thank my friends Arpan, Aruna, Jyothy, Lokendra (Loki), Raamesh, Sara, Harshad (Poddy), Saurabh (Spidey) and Vinay for their unconditional love and support over the years. I want to thank my wonderful parents-in-law for their blessings. My sincerest gratitude to my brothers, my grandmother and my parents for taking care of me, lending me emotional support, and constant praying for me. I owe you my life.

Last but not the least, I want to thank my husband, Varoon, whose love, support and encouragement helped me realize my potential.

Dedication

This thesis is dedicated to
my loving friend and husband, Varoon,
whose encouragement and support has brought me here.

Abstract

Recombinant protein therapeutics have transformed healthcare by paving the way for the treatment of refractory illnesses like cancer and arthritis. Chinese hamster ovary (CHO) cells are the major workhorse for the production of these therapeutics. Striving for continual improvements in the productivity and quality of protein produced in CHO cells, many process enhancements have been successfully implemented. However, many processes are still empirical, and we have little understanding of the mechanisms for these methods.

The availability of genomic resources for CHO cells has ushered in a 'genomics' era in bioprocessing. Genomic resources can now be employed to understand and improve cell lines and processes to enhance the productivity and quality of protein therapeutics produced by CHO cells. Seeking the development of genomic resources for CHO cells, the Chinese hamster genome and transcriptome were sequenced, assembled and annotated.

Such transcriptomic resources can be used to study the inherent transcriptomic variability in CHO cells. The genetic cues identified from the study of the variability in the glycosylation pathway genes opens up several opportunities to manipulate protein quality. The relative expression of isozymes in CHO cells affect metabolic characteristics, which in turn may potentially impact product quality or even process robustness. The comparative study of isozymes can give important clues for cell engineering and process development. The isozyme distribution in CHO cells indicates a very high overall glycolytic rate, insinuating to the possibility of manipulating glycolytic flux for improving processes.

Engineering superior metabolism through cell engineering can be used to reduce glycolytic flux in the late stage of the fed batch culture to reduce lactate accumulation. A novel dynamic promoter was used to drive the expression of a fructose transporter selectively in the late stages of the culture. By maintaining adequately low fructose levels in the late stage, the glycolytic flux was reduced significantly to induce lactate consumption. Since lactate accumulation is well accepted to be detrimental to productivity, this phenotype is desired for bioprocessing.

In addition to such high productivity processes, high producing cells are also desired. The lengthy process of cell line development transforms non-producing cells to high producers. The molecular changes in this transformation were elucidated by studying the transcriptome of CHO cells during cell line development. We hypothesize that methotrexate treatment not only increases the transgene copy number, but also enriches cells with superior growth, energy metabolism, and secretion capabilities. This leads to an enriched population of high producers. The sustenance of high productivity over several generations depends on the stability of the integration site of the transgene. Two methods for identifying the cell's transgene integration site were developed and optimized. These methods can be applied for high throughput investigation of stability of integration sites.

The application of genomics in bioprocessing has sparked a systems approach to investigate genetic regulation. This knowledge paved the way for controlling cellular metabolism and achieve stable and high producing cell lines and processes. Such genome scale analyses have a great potential to advance the capacity of CHO cells for biopharmaceutical applications.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

In the year 1982, Humulin, recombinant human insulin, produced in *E. coli* became the first recombinant protein to be licensed by the regulatory bodies for therapeutic applications. While *E. coli* can produce small protein molecules very effectively, they lack the necessary machinery to process larger and more complex molecules, especially glycoproteins like immunoglobulin G (IgG) or blood coagulation factor VIII. The production of such molecules requires post translational modifications such as glycosylation which is carried out in the endoplasmic reticulum (ER) and Golgi organelles in mammalian cells. Because of this, recombinant protein production soon resorted to mammalian cell as hosts for production.

The first mammalian cell based recombinant therapeutic, Activase by Genentech, was approved to be marketed by the US Food and drug Drug Administration (FDA) in the year 1987. Since then, the US FDA has approved over 100 recombinant protein therapeutics, culminating into global sales of US$120 billion in 2012 (Bandaranayake & Almo, 2014; Zhu, 2012).

Chinese hamster ovary (CHO) cells are indisputably the most popular cell line host for recombinant protein therapeutic production producing almost 60% of the therapeutics produced totally in mammalian cells. The ease of adaptation to suspension, serum-free or even chemically defined media makes them attractive for commercial production. The ability to use animal component-free media alleviates the risk of potential viral contamination coming from animal sources. These attributes have made CHO cells a major workhorse in bioprocessing.

The enormous success of these drugs in treating serious illnesses like cancer and arthritis has propelled drug discovery research to churn out an ever increasing number of drug candidates. All these candidates are very diverse in their structure, modality and requirement for post-translational modifications. This poses a great challenge for process engineers, who are required to develop robust processes for a large and diverse range of drugs. In the event of the patent expiration of older generation of therapeutics, follow-on

biologics, called 'biosimilars' are entering the market creating a wave of generic drugs. In order to cope with the increasing number of drug candidates, it has become imperative to expedite the development of these processes. However, since most stages in process development are largely empirical, the extensive optimization steps makes process intensification very difficult. To overcome this shortcoming, information gained from the knowledge of process variables influences on the host cells and products produced will prove to be valuable. A better understanding the process can help us make rational improvements.

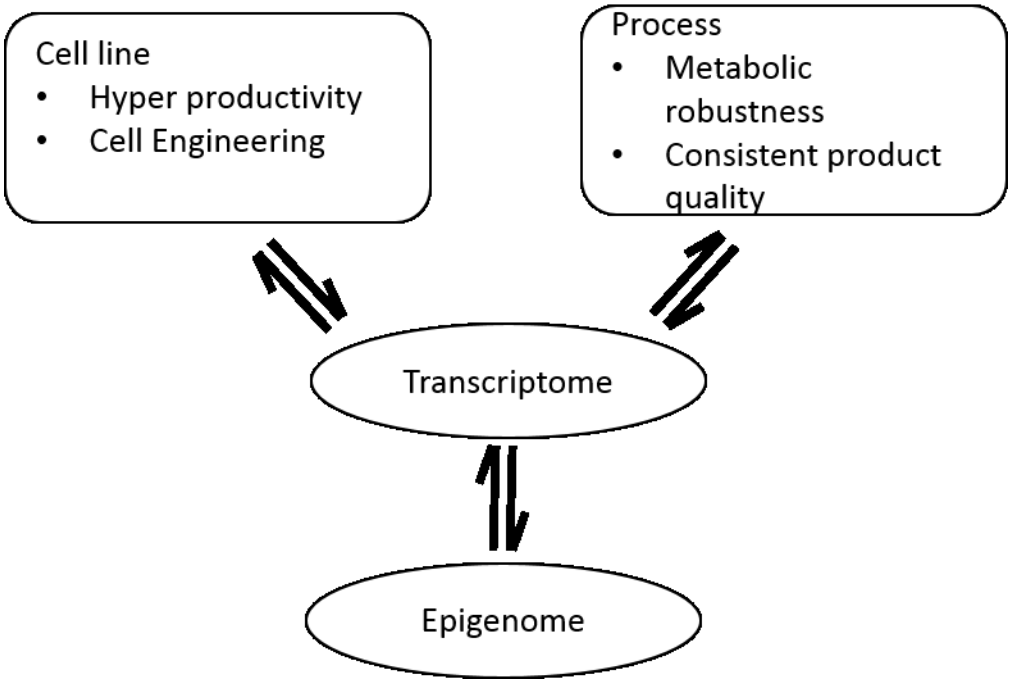


Figure 0-1: Role of genomics in bioprocessing

In the course of characterizing the impact of process variables, the cells cannot be treated as a black box as cellular physiology plays an important role in determining the product quality. Systems biology approaches are powerful tools that can be used to understand the molecular basis of changes in cellular behavior. Several thousands of variables can be measured in a single experiment. For example, using transcriptomics one can measure the gene expression levels of several thousands of genes in a single experiment. The use of systems biology was restricted in the past decade because of

insufficient genomic resources for CHO cells. The availability of the Chinese hamster genome (Brinkrolf et al, 2013; Lewis et al, 2013) and CHO cell genome (Xu et al, 2011) in the past two years, has revolutionized bioprocessing with the ability to harness genomics and transcriptomics methods for bioprocess advancements. The past three years have seen an overwhelming response to the use of genomics in CHO cell culture (reviewed in (Kildegaard et al, 2013)). Transcriptomics has helped to improve bioprocessing by providing us with scientific understanding of the hyper-productivity trait. Novel cell engineering strategies have improved product quality and yield (Figure 0-1).

Studies of the interaction of transcriptome with the epigenome are just beginning to emerge (Figure 0-1). The increasing evidence of epigenetic dysregulation in cancer cells makes us wonder about its prevalence and effect on the phenotype of CHO cells. Often we have attempted to seek master regulators for improving CHO productivity. Could this master regulator be an epigenetic event? The answer to this question can be found only through careful investigation of the epigenetic state of CHO cells.

This dissertation will focus on the application of genomic tools to understand and manipulate some physiological characteristics of CHO cells in the context of bioprocessing.

1.1 Thesis Organization

This thesis is an attempt to understand the molecular characteristics CHO cells in their role as recombinant protein producers. A brief background of mammalian cell culture along with a brief description of host cell lines that are commonly used in bioprocessing will be discussed in Chapter 2. A typical cell line development process is explained followed by a brief description of RNA sequencing and microarray technology for transcriptomics in Chapter 2. Chapter 3 summarizes the recent findings from transcriptome studies in CHO cells, and further discusses the challenges and the potential scope for further improvement. The transcriptomic diversity among different CHO host cells and producing cell lines derived from them is discussed in Chapter 4. A study characterizing the physiological changes in CHO cells during the cell line development process is

presented in Chapter 5. Chapter 6 shows an application of transcriptomic analyses in CHO cells to identify endogenous dynamic promoters. Such a promoter was used to engineer dynamic nutrient uptake in CHO cells leading to process improvements. Chapter 7 discusses the efforts towards building high quality genomic resources for CHO cells, including the assembly and annotation of the Chinese hamster genome and transcriptome. The development of methods for probing the transgene integration site in CHO cells is also described. Chapter 8 presents a brief conclusion and possible avenues for the future.

Chapter 2: Background

The first half of this chapter provides brief background in mammalian cell culture based production of recombinant protein therapeutics. Most of the background is given in the context of CHO cells because they are the most prominent hosts in biopharmaceutical production, and is also the focus of this dissertation. A description of the popular CHO host cell lines along with the details of their historical derivation is provided. This is followed by a description of cell line development process for selection of a stable CHO cell clone for production. The section on processes for protein production gives a brief summary of the fed-batch, perfusion and other processes for therapeutics production in mammalian cells. The second half of this chapter provides some background on the available genomic resources for CHO cells. This is followed by a brief description of the methods for transcriptome analysis, specifically RNA-sequencing and microarray technologies.

2.1 Recombinant Protein Production in Mammalian Cells

Mammalian cells produce almost half of all the currently approved biologics (Kantardjieff & Zhou, 2014). Their ability to fold many complex protein structures and to perform important post-translational modification like glycosylation make them essential for protein therapeutics production (Wurm, 2004). The glycan profile influences several characteristics of the protein drug, for example efficacy in eliciting immune response or the circulatory half-life. (reviewed in (Hossler, 2012; Hossler et al, 2009)).

2.1.1 Host Cells for Recombinant Protein Production

Protein therapeutics are commonly produced in human cell lines like HEK 29, 3T3, HeLa, HepG2 or rodent cell lines like mouse myeloma (NS0 and SP2/0), baby hamster kidney (BHK) or Chinese hamster ovary (CHO) cell lines, among the mammalian cell producers. The most prominent of these are CHO cells because of their ease of adaptability to different culture conditions, compatibility with large scale production, insusceptibility to human viruses and production of human compatible glycoforms. CHO cells alone

produce almost 60% of the mammalian cell based recombinant protein therapeutics and 30% of the currently approved global .biologics (Kantardjieff & Zhou, 2014).

Chinese hamster ovary (CHO) cell lines were derived from Chinese hamster ovary tissue biopsy over five decades ago, to study the karyotype of cells in tissue culture (Puck et al, 1958; Tijo & Puck, 1958). CHO cells have recently been given a "quasispecies" status because it is not a single cell line but a group of closely related cell lines having slightly different genomic compositions (Wurm, 2013). Since their derivation, they were among the most important cell lines used in biomedical research for decades. The low diploid number of chromosomes (2n = 22) in the Chinese hamster and functional haploidy made it relatively easy to derive mutant lines and hence, made these cell lines a valuable cytogenetic tool. They are also used as model organisms to study diabetes mellitus and toxicity studies. In the past two decades, CHO has also become the most prominent host cell for producing therapeutic proteins, acquiring much genomic reorganization in the process.

Cells that are isolated from tissue can normally grow only for a limited number of doublings after which they reach a 'crisis' state beyond which they no longer multiply. This is called the 'Hayflick phenomenon'. However, a few cells can overcome this state and become immortalized. CHO cells, are such immortalized cell lines. After its first derivation in 1958, the CHO cell line CHO-K1 was extensively distributed among different laboratories around the world for different cytogenetic studies. A few laboratories isolated mutants of CHO-K1 deficient in dihydrofolate reductase (DHFR) activity. CHO-DXB11 was derived in 1980 (Urlaub & Chasin, 1980), followed by CHO-DG44 in 1983 (Urlaub et al, 1983; Urlaub et al, 1986). The parentage details of only a few cell lines are known. Figure 0-2 shows the most accepted family tree of CHO cell lines. Several different parental CHO cells were further adapted to serum free or suspension medium at various laboratories.

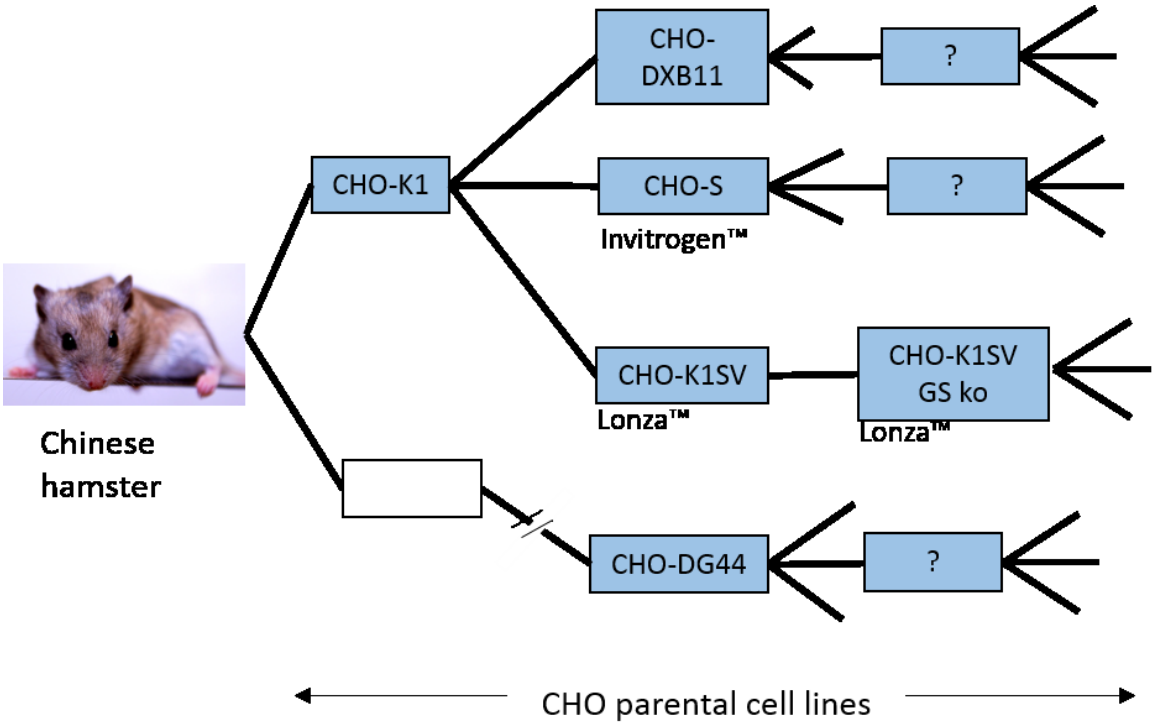


Figure 0-2: Best known genealogy of commonly used CHO host cell lines showing that the historical propagation of these different CHO host cell lines are rather different (Reconstructed from (Wurm & Hacker, 2011)).

The host cells that are available today may have genetically diversified due to repeated mutagenesis, adaptation and sub-culturing in the past five decades. Although the general phenotypes of these cells appear to be similar in the context of bioprocessing, their genotypes can be rather different. The differences in genotypes have been ignored in the past 25 years, potentially because of the unavailability of genomic resources. Also, probably there was no pressing need for understanding this diversity due to the limited range of molecules produced in CHO cells. With the increasing number of drug candidates, and the burst of genomic information available for CHO cells, it is now relevant to revisit this aspect to catalogue the genomic differences between CHO cell lines.

2.1.2 Cell Line Development

The transformation of CHO host cells into production cell lines begins with the introduction of the recombinant gene of interest along with a selection marker gene. In

most cases, the cells are also subjected to an amplification step after selection. These cells may or may not be sub-cloned using limited dilution at this stage.

Commonly used CHO host cell lines, such as DG44 and DXB11 (Figure 0-2), are auxotrophic mutants of the di hydrofolate reductase gene (dhfr) negative. Since they do not possess the dhfr gene, they cannot convert di hydrofolate into tetra hydrofolate. Tetra hydrofolate is necessary for DNA synthesis, and CHO cells deficient in DHFR activity require supplementation of hypoxanthine and thymidine (HT-media) for growth. By co-introducing the dhfr gene with the recombinant gene of interest and the selection marker, additional selection pressure can be applied by adding a dhfr antagonist. Methotrexate (MTX), a commonly used dhfr antagonist, is added to the cell culture medium. The concentration of MTX maybe gradually increased over a period of 15 days to intensify the amplification process. In order to cope with the high amount of selection pressure, the cells significantly amplify the dhfr genes, co-amplifying the surrounding genomic region including the recombinant gene. The selected cells typically have multiple copies of an approximately 100 kilo base of the chromosomal region containing the DHFR gene locus (Milbrandt et al, 1981). This process is called amplification because it results in a very high copy number, from hundreds to thousands of .copies of the transfected vector (Barsoum, 1990; Johnston et al, 1983; Kaufman & Schimke, 1981; Kaufman & Sharp, 1982; Kaufman et al, 1985; Ma et al, 1993; Mariani & Schimke, 1983; Niwa et al, 1991; Nunberg et al, 1978; Trask & Hamlin, 1989).

After MTX amplification, the cells are heterogeneous, so they have to be sub-cloned using limited dilution before mass production (Kim et al, 2001; Kim et al, 1998b). A few sub-clones are selected based on their growth and productivity, and later expanded. The cells are usually adapted to suspension culture in order to facilitate their cultivation in bioreactor vessels. Their performance is evaluated in shake flasks, and later in bioreactors. A few clones are selected and banked (Figure 0-3).

To test for stability, the selected clones are propagated for multiple generations through which the growth and productivity of the cells is monitored. The clone showing stable expression is selected for the final production.

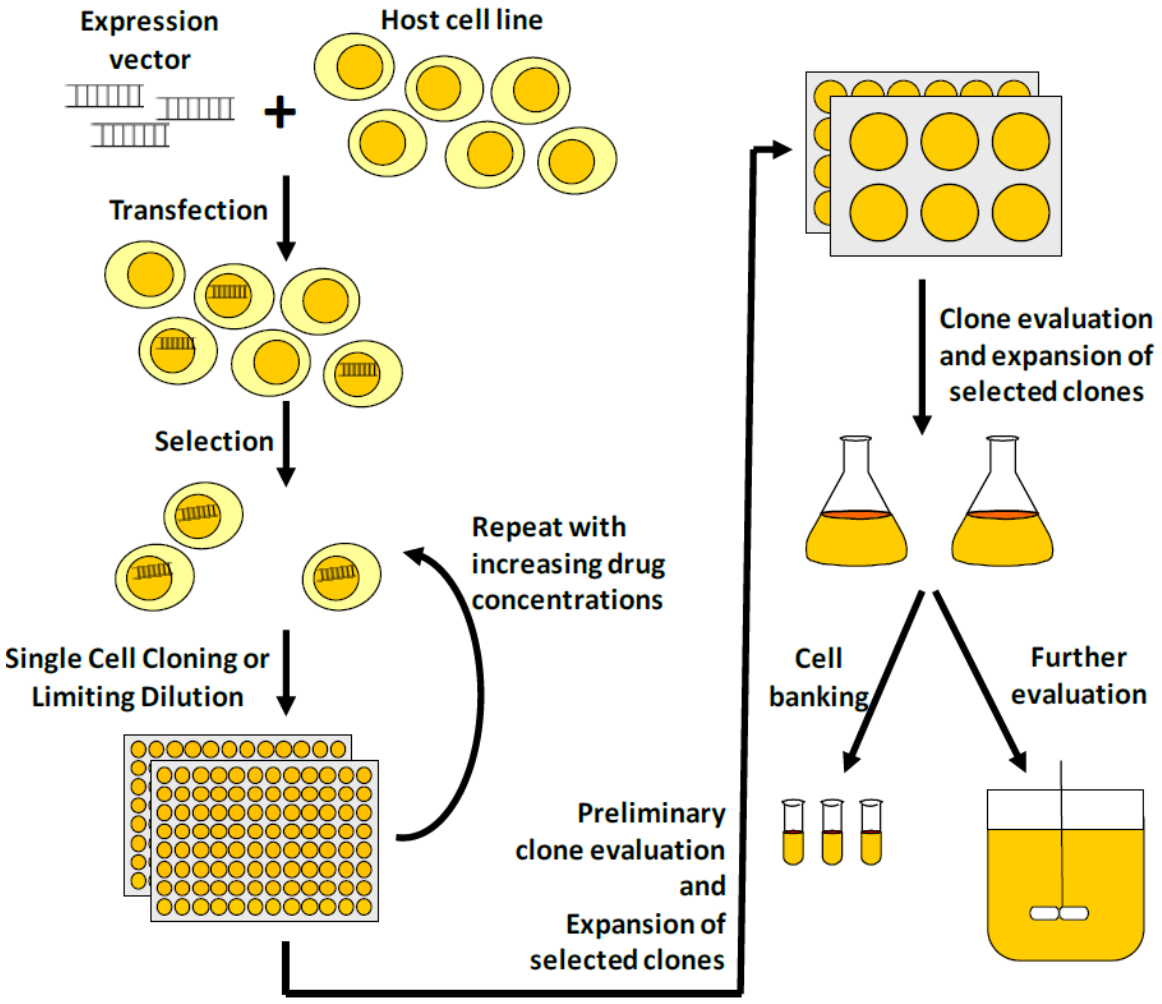


Figure 0-3: A typical cell line development process for recombinant protein production in mammalian cells. The host cell is first transfected with the recombinant gene along with a selection marker in order to select for successfully transfected cells. If amplification systems are used, the cells are subjected to high concentrations of the drug (MTX or MSX). Following amplification, cells are sub-cloned in limited dilution and screened for clones producing high amount of the recombinant product. A few candidate clones may be adapted to suspension culture and banked. In parallel, the candidate clones are characterized for stable production in shake flasks for many generations after which the most suitable clone is selected for production (Adapted from (Lai et al, 2013)).

The process of cell line development typically requires about 6 – 12 months, and is labor and capital intensive (reviewed in (Jayapal et al, 2007)). In past few years, there have been many advancements in this field towards increasing automation and process

reproducibility (reviewed in (Lai et al, 2013)). On the other hand, there has been little progress in understanding the molecular changes within CHO cells during cell line development.

The cell line development process has afforded the transformation of the host cells which secrete almost no protein, to become professional secretors. The cells produce almost 50 pg/cell/day which is equivalent to the productivity of high secretors in the human body like plasma cells or liver cells. Even after three decades following the development of this method, we do not fully understand the mechanisms of this transformation of non-producers into hyper-producers.

With the increasing number of drug candidates, especially in past few years, it has become necessary to accelerate the time required to develop a cell line. A better understanding of the transformation can enable us not only to speed up the process of cell line development, but also to develop more rational methods to screen cells with hyper productivity traits.

2.1.3 Processes for Recombinant Protein Production

The stable clone selected after the cell line development process is typically grown in bioreactors under controlled environmental conditions of pH, temperature, dissolved oxygen and CO_2 for producing high amounts of therapeutic protein for clinical use. The cells are usually cultured in fed batch mode or perfusion mode. In fed batch mode, the cells are inoculated at a low concentration in the bioreactor and grown to high concentrations. Concentrated feed medium is fed regularly, typically once a day to make up for the consumed nutrients. The volume of the reactor increases through the course of the culture. To accommodate for the increasing volume, these cultures are usually initiated at volumes lower than the capacity of the bioreactor (reviewed in (Wlaschin & Hu, 2006)). The cells are grown to the maximum cell density after which cell growth is inhibited because of the accumulation of toxic metabolites like lactate and ammonia (Ozturk et al, 1991; Zhou et al, 1997). The high cell density is maintained until the viability drops to low levels because of high osmolality and reactive oxygen species (ROS). Since the protein quality is affected

at low viabilities of the culture (Yang & Butler, 2002), the culture is terminated at a certain predetermined viability, usually between 50% and 80%. The product is harvested at the end of the culture. A fed batch culture typically lasts for 14 days and routinely yields titers of up to 5 g/L (Whitford, 2006). The product titer largely depends on the specific productivity of the cells, the peak cell density, and the culture duration.

Some processes in fed batch bioreactors employ additional productivity enhancing conditions such as exposure to low temperature, or addition of sodium butyrate in the late stage of the culture after the cells reach maximum cell density. Other ways to increase productivity in the stationary phase are increasing osmolality or a pH shift (Baik et al, 2006; Birzele et al, 2010; De Leon Gatti et al, 2007; Fogolin et al, 2004; Jiang & Sharfstein, 2008; Kantardjieff et al, 2010a; Kaufmann et al, 1999; Klausing et al, 2011; Moore et al, 1997; Oh et al, 1993; Rodriguez et al, 2005; Shen et al, 2010; Shen & Sharfstein, 2006; Trummer et al, 2006; Yee et al, 2008; Yoon et al, 2004; Yoon et al, 2003a; Yoon et al, 2003b). Sodium butyrate is a histone deacetylase inhibitor (Candido et al, 1978; Davie, 2003; Sekhavat et al, 2007), and may play a role in increasing transgene accessibility in the stationary phase, leading to higher expression and productivity of the recombinant gene.

In contrast to fed batch modes, in continuous culture, the cells are cultivated in a bioreactor with the feed media being added continuously. Simultaneously, an equivalent volume of reactor slurry is withdrawn, keeping the reactor volume constant. The product is extracted from this efflux continuously. One of the disadvantages of continuous culture is that the flow rate is limited by the growth rate of the cells. Continuous cultures cannot be operated at higher rates than the cellular growth rate, or else the cells will be washed out. Overcoming this shortcoming, in perfusion culture, the cells are separated from the slurry by filtration and then recycled back from the retentate into the reactor (Abu-Absi et al, 2014). One of the disadvantages of continuous culture is that the flow rate is limited by the growth rate of the cells. Continuous cultures cannot be operated at higher rates than the cellular growth rate, or else the cells will be washed out. One way of overcoming this shortcoming, is the filtration of cells followed by recycling them to the reactor (Abu-Absi

et al, 2014). The cell concentration in the reactor remains constant at usually 60 to 100 million cells/mL. The product is harvested continuously from the filtrate. Since the spent media is continuous withdrawn from the reactor it prevents the buildup of toxic metabolites like lactate and ammonia. This allows the bioreactor to be run for much longer durations. Perfusion cultures can achieve 1-2 g/L/day titers (Bonham-Carter & Shevitz, 2011).

The fed batch and perfusion modes are by far the most commonly used processes for commercial production. Among them, though fed batch culture is the dominant industrial practice. The choice of production format also depends on characteristics of the product and the production scale. For example, for an unstable product, a perfusion process maybe more suited due to much lower residence time. The average residence time is in the order of hours in the perfusion process compared to days in a fed batch process (Kadouri & Spier, 1997).

2.2 Genomics and Transcriptomics

The number of genomes being sequenced around the world is exponentially increasing (Liolios et al, 2010) fuelled by the plummeting costs of high throughput sequencing. With the sequencing of the CHO-K1 and Chinese hamster genomes, genomic resources for CHO cells have piled up significantly over the past few years (Brinkrolf et al, 2013; Lewis et al, 2013; Xu et al, 2011). Genomics takes a holistic approach for the investigation of the cellular properties as opposed to the traditional individualistic approach of studying a small portion of the genome in detail. Genomic investigations generate large amounts of data containing a wealth of information. A major challenge lies in finding intelligent ways of sifting through these vast amounts of data to obtain relevant physiological information. The availability of genomic resources for Chinese hamster has facilitated the use of genomics to characterize the CHO cells.

Transcriptomics is the study of the expressed region of the cells' genome including mRNA, rRNA, tRNA and other non-coding RNA. The most widely studied form of transcriptome is the mRNA. In a transcriptomics study, the expression level of all the genes

within a cell is measured in a single experiment. The most common tools for transcriptome analysis are microarray or RNA-Sequencing.

2.2.1 Microarray

Microarrays are used for global profiling of gene expression. Oligonucleotide microarrays are the most common form of microarrays that have evolved from cDNA microarrays. Each transcriptome microarray chip contains several spots, where each spot is a probe for the gene expression of a gene.

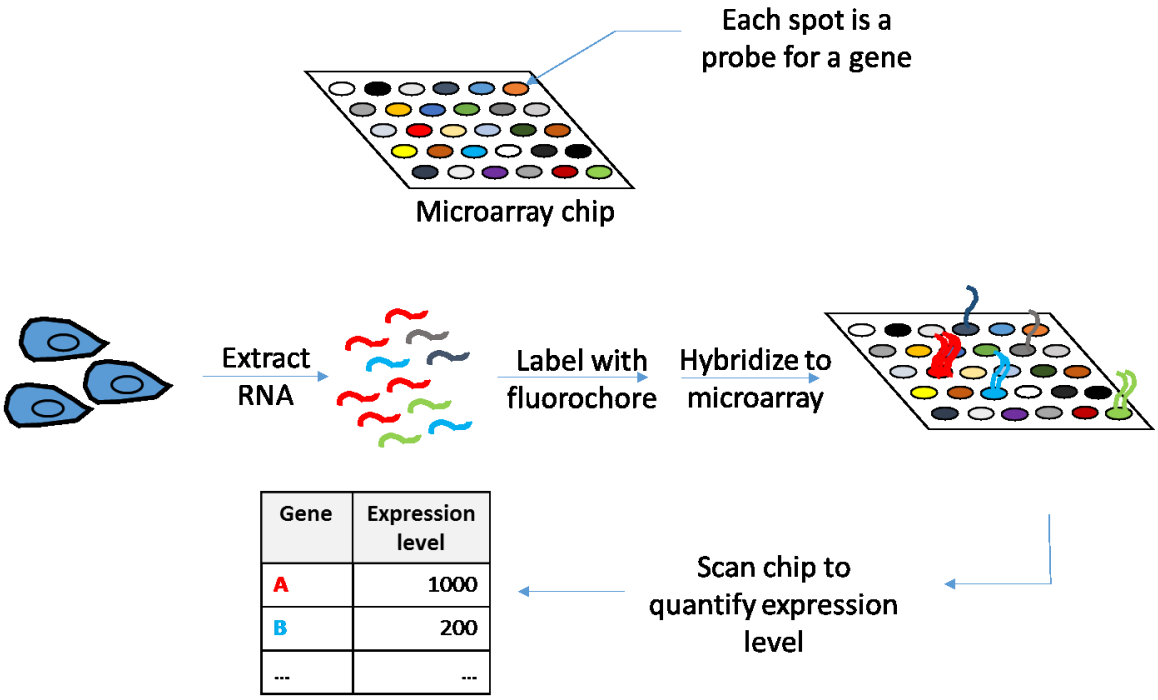


Figure 0-4: An illustration of the working principle of an oligonucleotide microarray

Usually, there are multiple probes per gene. The spot contains many copies of the reverse strand of a portion of the cDNA of the gene to be probed. The selective unique base pairing of nucleic acids of DNA is exploited to probe the expression levels of all the genes. Total RNA extracted from cells is converted to cDNA using oligo-dT primers for enriching mRNA from the sample. The cDNA is either directly hybridized to the microarray or converted to cRNA prior to hybridization. The cRNA nucleotides are labelled with

fluorochores to enable the measurement of gene expression level via fluorescence intensity. Spots on the chip are scanned to quantify the gene expression levels of all the genes in a single experiment (Figure 0-4).

To take advantage of the microarray technology, it is imperative to have a comprehensive coverage of the transcriptome along with a good quality of annotation for most of the genes. With newly sequenced draft genomes, it is difficult to secure such high quality annotation. Over the years, our lab has invested a lot of resources in expanding the genome and transcriptome sequence annotation for Chinese hamster and CHO cells (Jacob et al, 2010; Kantardjieff et al, 2009; Wlaschin & Hu, 2007b; Wlaschin et al, 2005). The lab has designed, constructed and validated several different microarrays for CHO cells, a few of which will be referred to in this thesis.

Several different commercial platforms are available for constructing oligonucleotide microarrays. While the working principle is the same, the manufacturing technologies and feature properties may vary quite a bit across different platforms. Each probe maybe 25 bp or 60 bp in length depending on the platform used. In our lab, we have designed and validated microarrays in both Affymetrix (25 bp probes) and Nimblegen (60 bp probes) platforms. In this dissertation, I have used the Nimblegen microarray for the work described in Chapter 4 and the Affymetrix microarray for the work described in Chapter 5.

Microarrays can also be used to probe genome copy number variation. These microarrays usually have probes either tiling the entire genome, or evenly spaced across the genome where the copy number of the genomic region is probed. Such arrays are called comparative genomic hybridization (CGH) arrays.

For transcriptome analysis, our lab has generated several generations of microarrays, continuously validation.

2.2.2 RNA-Sequencing

RNA sequencing based transcriptome quantification methods is a powerful method that has gained acceptance and popularity in the past few years. Often referred to as RNA-

Seq, it involves direct sequencing of RNA using high throughput sequencing methods. The short sequencing reads are aligned to the reference genome or transcriptome. The number of reads mapping to a gene is a measure of the expression level of that gene (Figure 0-5). With the decreasing cost of high throughput sequencing, the costs of conducting an RNA-Seq experiment has reduced leading to its increase in popularity.

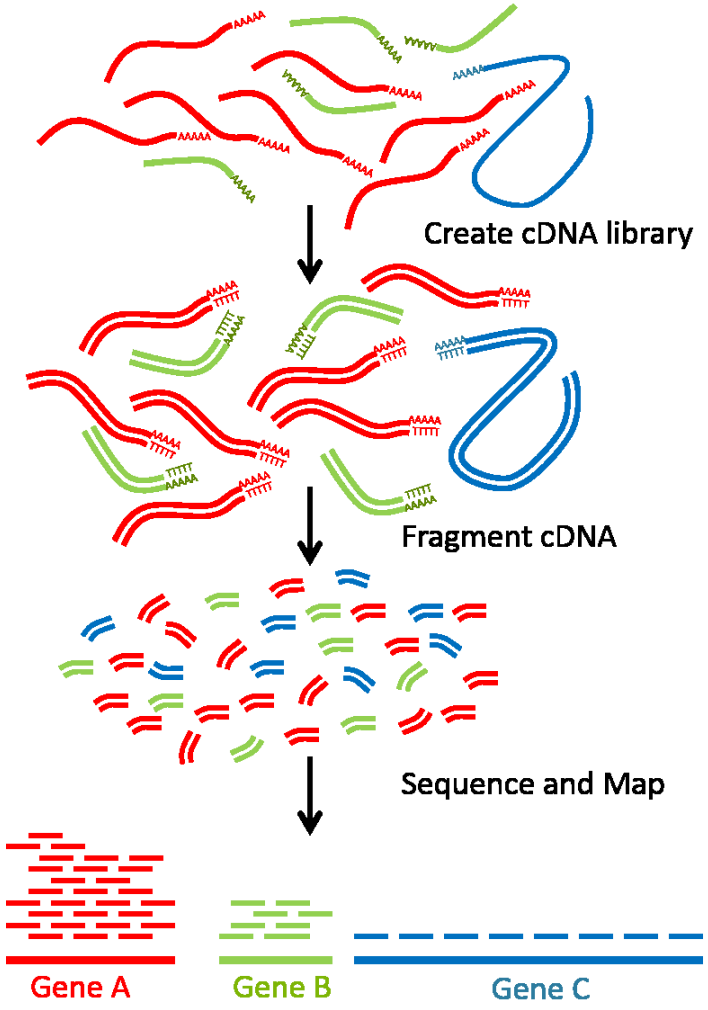


Figure 0-5: Transcriptome analysis by RNA-Seq. The total RNA is converted to cDNA which is fragmented and sequenced. The sequencing reads are mapped to all the genes, and the depth of coverage of the genes is used to quantigy gene expression level.

Similar to the microarray experimental procedure, the total RNA is extracted from the cells and converted to cDNA using oligo-dT primers, thereby enriching for mRNA from total RNA. The cDNA is fragmented into approximately 200 – 500 bp size fragments. The fragments are sequenced by high throughput sequencing. Each sequencing output reads out the first 50 – 100 bp of the fragment depending on the experimental design. This output is called a read. The read sequences are usually preprocessed to remove low quality sequence and adaptor sequences after which they are mapped to the reference genome or transcriptome. Mapping is the process of identifying the genomic source of the read by simply matching or aligning the reads to the reference. As shown in Figure 0-5, the gene expression is quantified from the depth gene coverage.

In addition to transcript quantification, RNA-Seq can provide additional information of novel transcript expression, or unique alternatively spliced forms, or even single nucleotide variants.

Chapter 3: Transcriptomics in Bioprocess Development

3.1 Summary

Global survey of transcriptome dynamics can provide molecular insights into cell physiology. In the past few years, DNA microarray for transcriptome analysis has been augmented by high-throughput sequencing methods, extending the reach of transcriptome analysis to species of biotechnological importance, for which the development of genomic tools has been lagging. The rapid accumulation of sequencing data for these species highlighted the need of more evidence-based annotation. Recent findings in epigenetic regulation in human and mouse will inspire similar research in CHO and BHK cells. Transcriptome studies in these recombinant cells will likely lay the foundation for systems based genome engineering for develop superior producing cell lines. Herein, we summarized recent findings and advances in transcriptome studies of cell culture bioprocesses. The potential impact of transcriptomics on biopharmaceutical process technology is also discussed.

3.2 Introduction

Transcriptomics is a powerful tool providing a global view of cellular biological activities. Tens of thousands of different variables are simultaneously measured. Even proteomics or metabolomics cannot match the magnitude of quantitative information that can be obtained from a single transcriptomic assay.

Recent advances in DNA sequencing technology have made global exploration at the transcriptome and genome levels feasible and affordable. This article reviews what have been accomplished in transcriptome analysis in the past few years and the possibility of its reach in the near future.

3.3 Transcriptomic Tools

Until a decade ago, global assessment of transcriptome was accomplished by sequencing expressed sequence tags (ESTs) or similar methods such as serial analysis of

gene expression (SAGE). Those decade long efforts in the 1990s and early 2000s collected a large number of EST sequences derived from different tissues in either healthy or disease states from various species. For human and mouse, over ten million of EST sequences were accumulated by 2006 (Nordstrom et al, 2006). Different ESTs with overlapping sequences were then assembled into contiguous sequences (contigs) or even full-length transcripts. More importantly, those early efforts in sequencing and assembling for human and mouse were accompanied by arduous efforts in annotation to give them the physiological context that paved the way for later genome annotation. The early work on ESTs demanded a large amount of resources (for a review on EST annotation, see (Kawai et al, 2001)). In ensuing sequencing effort in other mammalian species, including Chinese hamster and Syrian hamster used in biopharmaceutical production, the scale of sequence information and annotation effort is dwarfed by those in human and mouse.

Through EST sequencing, especially methods like SAGE, one can also assess the transcript abundance level through the frequency of E. coli clones (each clone contains one EST species) that was sequenced. However, the expenditure of such an effort prevented it from being used routinely. Synthetic oligo DNA arrays and cDNA microarrays made transcriptome analysis more accessible. However, these microarrays were possible only for species for which EST sequence information was available. Nevertheless, transcriptome analysis was soon applied to antibody producing cell lines to explore gene expression changes under different metabolic states in a continuous culture (Korke et al, 2004).

The versatility of expression microarrays spurred some early efforts to sequence ESTs of Chinese hamster tissues and CHO cells using the conventional Sanger sequencing method on phage libraries (Kantardjieff et al, 2009; Wlaschin et al, 2005). The development and the rapidly decreasing cost of high-throughput sequencing prompted its use to sequence transcripts of the industrially important CHO cell lines and Chinese hamster tissues (Becker et al, 2011; Birzele et al, 2010; Jacob et al, 2010; Rupp et al, 2014). However, compared to human or mouse, the number of Chinese hamster or CHO ESTs available is still very small. More recently, a transcriptome reference of Syrian hamster tissues and BHK cell lines (derived from Syrian hamster) was established using a

combination of Sanger and Illumina platforms (Johnson et al, 2013). Next-generation sequencing technologies also ushered the sequencing of the CHO-K1 cell line genome (Xu et al, 2011) and Chinese hamster genome (Lewis et al, 2013), and later, chromosome sorted Chinese hamster genome (Brinkrolf et al, 2013). An initial effort to sequence the Syrian hamster genome is currently underway (2013). From the genome sequence, *in silico* prediction of gene coding regions can be made.

While millions of ESTs are sequenced for mouse and human, only 860 and 12 ESTs are available for Syrian and Chinese hamster, respectively, in the NCBI dbEST database. The availability of EST sequences enabled the construction of expression microarray for transcriptome surveys from early custom cDNA (De Leon Gatti et al, 2007; Wong et al, 2006; Yee et al, 2009) to oligoDNA custom and commercial arrays (Jayapal & Goudar, 2014; Kantardjieff et al, 2009; Melville et al, 2011), although the depth of sequence coverage varied over a range in those arrays (Baik et al, 2006; Ernst et al, 2006; Trummer et al, 2008).

3.3.1 RNA-Seq

High-throughput sequencing of ESTs is not only used to assemble transcript coding sequence of genes but also to assay the abundance distribution in a transcriptome (RNA-Seq) even in a species without prior genome or transcriptome sequence information. These high-throughput sequencing methods generate sequences of short fragments of a transcript (called "reads") (Jacob et al, 2010). A general procedure is to assemble those reads into contigs, or map those reads into a previously assembled reference transcriptome. Since the transcript length varies widely, the number of reads mapped to each contig (or gene) is normalized to its length. The abundance level of different contigs is indicated by number of reads per unit contig length. Unlike fluorescence intensity-based expression arrays, for which the detection of rare transcript is limited by sensitivity of fluorescence detection and the quantification of highly abundant genes suffers from intensity saturation, RNA-Seq is relatively free from those biases. By sequencing to a great depth (i.e., increasing the total number of reads sequenced), RNA-Seq can reveal any transcriptome in greater detail over

a wider dynamic range than DNA microarray. Through RNA-Seq, the abundance level of transcripts in CHO cells was shown to vary up to of five orders of magnitude. As expected, the transcripts of heavy chain and light chain IgG genes in the high producing recombinant CHO cell line were among the most abundant transcripts (Jacob et al, 2010).

Along with providing a very high sequencing depth for abundantly expressed genes (Birzele et al, 2010), RNA-Seq also offers a high sensitivity in detecting sequence variants, i.e., more than one type of nucleotide is represented in the same position of a transcript (Johnson et al, 2013). These sequence variants may be caused by heterozogosity or by the accumulation of mutations. Interestingly, the frequency of variants with a high statistical confidence in CHO and BHK cell lines examined was not high, considering that mutations could possibly accumulate in each individual cell within the population. In spite of extensive culture of the cell lines, no sequence variant in the product transcript was detected.

3.3.2 Challenges

The ideal way of annotating a genome is based on evidence of transcription and biological functional studies. Among all mammals, human, mouse and to a much lesser extent, rat, have large EST repertoires. The data on mutation loci and genetic studies accumulated over years greatly facilitate evidence-based annotation for human and mouse genes. For example, the manually annotated mouse database of the FANTOM consortium (Kawai et al, 2001), the repertoire of the ENCODE project, (Bernstein et al, 2012) and the Cancer Atlas project for human (2008) are great resources for relating genes to their expression and functions.

For Syrian hamster and Chinese hamster, such data are meager or even nearly non-existent in comparison. Like all recently sequenced genomes, genes in the Chinese hamster genome have been identified through *de novo*, homology-based, and transcriptome-aided prediction (Lewis et al, 2013). The EST data can be assembled separately and may be annotated by homology to other species (Becker et al, 2011; Birzele et al, 2010; Rupp et al, 2014). Using next-generation sequencing technology as the primary means of

sequencing, the resulting EST assemblies are often fragmented contigs, except for very abundant genes. Many transcripts are not full-length sequences of the transcribed genes. In general, a reasonably large fraction (~30%) of genes can be annotated by homology with a high degree of confidence. But many are poorly annotated for lacking highly homologous orthologs in other better annotated species (Becker et al, 2011). RefSeq IDs for CHO-K1 genome are not yet accepted by functional analysis modules such as IPA, GSEA, or GenMAPP. To utilize the resources available for human and mouse, one currently still relies on tagging CHO and BHK genes to the Ensembl IDs of their human or mouse orthologs. The focus of transcriptome research in recombinant cells therefore needs to shift from sequencing to annotation in order to converge with those performed in human and mouse.

3.3.3 Data Analysis

Another challenge to transcriptome analysis in biopharmaceutical industry is that the degree of differential expression in the host and production cell lines is relatively small. Upon differentiation or developmental events, many genes change their expression by several orders of magnitude. In contrast, gene expression changes in cell lines under different culture conditions or different treatments are relatively moderate. Given that cells under somewhat different culture conditions may have different total RNA and mRNA contents and the inevitable experimental variations, the identification of differential expressed genes is easily affected by the method of normalization used in data analysis. It might be prudent to perform analysis on biological replicates to estimate the extent of systematic errors for genes expressed at different abundance levels to assist making differential expression calls.

RNA-Seq based transcriptome analysis will be increasingly used because of its independence from a prefabricated expression array. RNA-Seq data, represented by discrete read count, often follow a Poisson, binomial, or multinomial distribution (Fang et al, 2012). Because of this drastic difference in data nature, normalization and statistical analysis methods are different from conventional DNA microarrays. Instead of using t-test

or ANOVA, which are based on normal distribution assumption, it is more appropriate to apply Fisher's exact test with or without an estimation of the conditional maximum likelihood. RNA-Seq normalization methods take sequencing depth and transcript length into consideration to account for intrinsic variance prior to analysis (reviewed in (Dillies et al, 2013)). Among them, fragments per kilobase per million mapped reads (FPKM) are commonly used metric. Recently, a new normalization method based on expected uniquely mappable area of genes or isoform models (NEUMA) has been shown to produce better consistency with qRT-PCR compared to other methods (Lee et al, 2011).

Furthermore, like other biochemical variables, transcript levels change dynamically in bioprocesses. Instead of static comparison, or finding genes whose transcript levels change at the same time point, it is more revealing to identify genes whose time profiles are altered under different culture conditions. Because cultures are not synchronized, time alignment may be necessary before dynamic differential expression call, hierarchical clustering, or pathway analysis can be performed (for review see (Castro-Melchor et al, 2011)).

3.4 Transcriptome Dynamics and Regulation in Cell Line Development and Bioprocessing

3.4.1 Transcriptome Dynamics in Recombinant Cell Culture

Because of the limited availability of microarrays for species of cell lines used in biopharmaceutical industry except for human and mouse, and perhaps also the cost associated with microarray assays, the research literature on transcriptome analysis in bioprocessing of biologics is still relatively thin and mostly from a few research groups. As expected, the early work often addressed the issues of productivity of recombinant proteins by comparing cell lines of different productivities or culture conditions that elicit different productivities. A common thread in the findings from those studies was that the degree of differential expression among samples was small and the number of replicates needed to obtain a low false discovery rate was often too large to be affordable. Instead of

focusing on differential expression at individual gene levels, gene functional class analysis was performed. These methods use pathway analysis tools to identify groups of functionally related genes with statistically significant changes. Even though individual genes may incur only relatively minute changes in expression, if a large number of such genes in a cellular functional class (or gene set) incur such changes, the functional class can be identified as differentially expressed by those methods.

Gene functional class analysis on transcriptome data upon sodium butyrate treatment and/or temperature shift (Baik et al, 2006; Birzele et al, 2010; De Leon Gatti et al, 2007; Kantardjieff et al, 2010a; Klausing et al, 2011; Wippermann et al, 2014; Yee et al, 2008; Yee et al, 2009) revealed that gene sets involved in protein processing and secretion, signaling pathways, and cell cycle were differentially expressed. Similar pathways are observed to be altered in high and low producer comparison (Clarke et al, 2011; Nissom et al, 2006; Seth et al, 2007b; Vishwanathan et al, 2013), methotrexate (MTX) treatment for gene amplification (Grillari et al, 2001) and productivity-enhancing conditions such as osmotic stress (Shen et al, 2010; Shen & Sharfstein, 2006). Instead of having one or a small number of master regulator(s) differentially expressed, a set of genes in a few functional class were found to alter their transcript levels.

Transcriptomic data have been exploited to identify targets for cell engineering. From CHO microarray data, genes differentially expressed with temperature were identified and a representative promoter region isolated (Thaisuchat et al, 2011). This promoter was shown to induce expression of a luciferase reporter gene upon temperature shift. The system may be used to control the expression of a toxic protein or to induce anti-apoptosis genes. In a similar vein, endogenous CHO promoters which drive transcripts in different dynamic patterns throughout fed-batch cultures were identified. One of them, the promoter of Thioredoxin-interacting protein (Txnip) which has a lower expression level in the exponential growth stage but a high level in the late stage, was used to express a sugar transporter in a dynamic fashion to manipulate glycolytic flux in the late stage (Le et al, 2013).

3.4.2 Transcriptomics and Hyper-productivity in Cell Line Development

A recent study examined changes in transcriptome during cell line development using methotrexate-induced amplification of dihydrofolate reductase (DHFR) gene (Vishwanathan et al, 2013). An interesting finding was that the transcripts of product genes (the heavy chain and light chain of immunoglobulin G) were already at high levels upon the selection of cells bearing the vector and increased little after amplification. This is contrary to the conventional notion that amplification serves to increase the copy number of the transgene and its transcript level. It was hypothesized that the amplification step may select for surviving cells that have developed cellular machineries to cope with the surge of secretory protein synthesis. This transcriptome data set was combined with previous meta-data compiled from comparison of high and low producers and high- and low-productivity culture conditions (Charaniya et al, 2009; De Leon Gatti et al, 2007; Kantardjieff et al, 2010a; Seth et al, 2007b; Yee et al, 2008; Yee et al, 2009) in an attempt to generate a hyper-productivity gene set. The majority of these genes are involved primarily in protein synthesis, metabolism, cell cycle, and transcription regulation. With the extremely large number of dimensions, and with the possibility of alternative routes to give rise to the hyper-productivity trait, the development of a hyper-productivity gene set will require a very large set of data of cell lines and processes of different productivities, and will likely be accomplished only through a large scale and collaborative effort (Figure 0-6 and Table 0-1).

Table 0-1: Pathways correlated to the hyper-productivity trait compiled from studies studying the hyper productivity trait in different aspects of bioprocessing mentioned in Figure 0-6.

Pathways affected	Temperature Shift (TS)	Sodium butyrate treatment (SB)	TS and SB	Osmotic shock	Cell line development-selection	Cell line development - amplification	High versus Low producer comparison
Cell cycle	(Barron et al, 2011)	(Birzele et al, 2010; De Leon Gatti et al, 2007; Klausing et al, 2011; Yee et al, 2008)	(Kantardjieff et al, 2010a)	(Shen et al, 2010; Shen & Sharfstein, 2006)	(Vishwanathan et al, 2013)	(Vishwanathan et al, 2013)	(Charaniya et al, 2009; Nissom et al, 2006; Seth et al, 2007b; Trummer et al, 2008; Vishwanathan et al, 2013)
Cell proliferation	(Barron et al, 2011)	(Birzele et al, 2010)		(Shen & Sharfstein, 2006)			(Nissom et al, 2006; Seth et al, 2007b)
Cell adhesion	(Barron et al, 2011)						
Apoptosis	(Barron et al, 2011; Wippermann et al, 2014)	(De Leon Gatti et al, 2007; Klausing et al, 2011; Yee et al, 2008)		(Shen et al, 2010; Shen & Sharfstein, 2006)			(Seth et al, 2007b)
Lipid, fatty acid, steroid metabolism	(Barron et al, 2011)	(De Leon Gatti et al, 2007)	(Kantardjieff et al, 2010a)	(Shen & Sharfstein, 2006)			(Clarke et al, 2011; Trummer et al, 2008)
Protein synthesis, processing, trafficking and secretion	(Baik et al, 2006; Yee et al, 2009)	(De Leon Gatti et al, 2007; Yee et al, 2008)	(Kantardjieff et al, 2010a)	(Shen et al, 2010; Shen & Sharfstein, 2006)	(Vishwanathan et al, 2013)	(Vishwanathan et al, 2013)	(Charaniya et al, 2009; Clarke et al, 2011; Seth et al, 2007b; Trummer et al, 2008; Vishwanathan et al, 2013)

Pathways affected	Temperature Shift (TS)	Sodium butyrate treatment (SB)	TS and SB	Osmotic shock	Cell line development-selection	Cell line development - amplification	High versus Low producer comparison
Proteolysis	(Yee et al, 2009)					(Vishwanathan et al, 2013)	(Charaniya et al, 2009)
DNA replication		(Birzele et al, 2010)			(Vishwanathan et al, 2013)	(Vishwanathan et al, 2013)	(Seth et al, 2007b; Vishwanathan et al, 2013)
Oxidative stress	(Yee et al, 2009)	(De Leon Gatti et al, 2007; Yee et al, 2008)					(Trummer et al, 2008)
Cellular assembly and organization	(Yee et al, 2009)	(Birzele et al, 2010)	(Kantardjieff et al, 2010a)	(Shen & Sharfstein, 2006)		(Vishwanathan et al, 2013)	(Charaniya et al, 2009; Seth et al, 2007b; Trummer et al, 2008; Vishwanathan et al, 2013)
DNA Repair	(Wippermann et al, 2014)			(Shen & Sharfstein, 2006)			(Seth et al, 2007b)
Nucleotide metabolism	(Barron et al, 2011)	(De Leon Gatti et al, 2007)		(Shen & Sharfstein, 2006)			(Nissom et al, 2006)
Transcription and Translation	(Wippermann et al, 2014)	(De Leon Gatti et al, 2007; Klausing et al, 2011)	(Kantardjieff et al, 2010a)	(Shen et al, 2010; Shen & Sharfstein, 2006)			(Charaniya et al, 2009; Clarke et al, 2011; Grillari et al, 2001; Nissom et al, 2006; Seth et al, 2007b; Trummer et al, 2008)

Pathways affected	Temperature Shift (TS)	Sodium butyrate treatment (SB)	TS and SB	Osmotic shock	Cell line development-selection	Cell line development - amplification	High versus Low producer comparison
Histones		(De Leon Gatti et al, 2007)					(Charaniya et al, 2009; Nissom et al, 2006)
Signaling	(Baik et al, 2006)		(Kantardjieff et al, 2010a)	(Shen et al, 2010; Shen & Sharfstein, 2006)	(Vishwanathan et al, 2013)	(Vishwanathan et al, 2013)	(Charaniya et al, 2009; Nissom et al, 2006; Seth et al, 2007b; Vishwanathan et al, 2013)
Growth factor signaling			(Kantardjieff et al, 2010a)		(Vishwanathan et al, 2013)	(Vishwanathan et al, 2013)	
MAPK signaling	(Wippermann et al, 2014)						
Wnt β-catenin signaling	(Wippermann et al, 2014)						
TGF-β signaling pathway		(Birzele et al, 2010)					
Golgi			(Kantardjieff et al, 2010a)				(Charaniya et al, 2009; Clarke et al, 2011)
Carbohydrate Metabolism*	(Baik et al, 2006; Barron et al, 2011; Yee et al, 2009)	(De Leon Gatti et al, 2007; Yee et al, 2008)		(Shen et al, 2010; Shen & Sharfstein, 2006)			(Nissom et al, 2006; Seth et al, 2007b; Trummer et al, 2008; Vishwanathan et al, 2013)

Pathways affected	Temperature Shift (TS)	Sodium butyrate treatment (SB)	TS and SB	Osmotic shock	Cell line development-selection	Cell line development - amplification	High versus Low producer comparison
Ion transport*	(Yee et al, 2009)	(Birzele et al, 2010; De Leon Gatti et al, 2007; Kantardjieff et al, 2010a)		(Shen et al, 2010; Shen & Sharfstein, 2006)			(Nissom et al, 2006; Seth et al, 2007b)
Amino acid or protein metabolism*	(Barron et al, 2011)	(Yee et al, 2008)		(Shen et al, 2010; Shen & Sharfstein, 2006)	(Vishwanathan et al, 2013)	(Vishwanathan et al, 2013)	(Grillari et al, 2001; Nissom et al, 2006; Trummer et al, 2008; Vishwanathan et al, 2013)

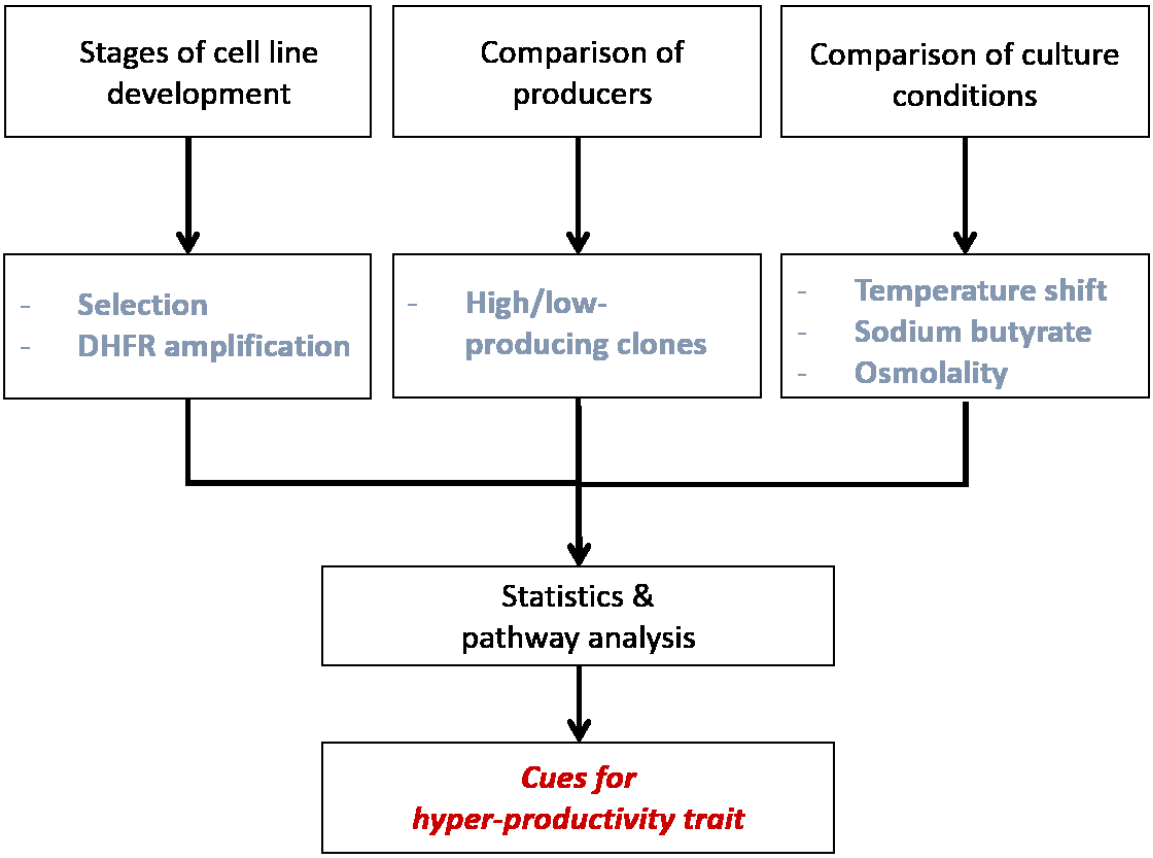


Figure 0-6: Towards generating a hyper-productivity gene set from meta-analysis of historical transcriptome data

3.4.3 Transcriptomics and Product Quality

The increasing accessibility of transcriptome survey has spurred interest in employing transcriptome analysis in help steering the glycosylation pattern of the product. The repertoire of glycosylation genes in mammals is large, but their expression is tissue specific. Among the 300 genes related to glycosylation identified in CHO genome, 159 are expressed in CHO K1 cell as revealed by transcriptome sequencing. Among those not expressed is α(2,6) sialyltransferase (Becker et al, 2011; Xu et al, 2011). Since α(2,6) sialic acid is present in many glycoproteins of human, de-silencing of this gene may lead to the synthesis of glycans closer to those of human origin, mitigating the need of cloning in α(2,6) sialyltransferase to accomplish the same glycan alteration. A sub-array of 79 glycosylation related genes was used to study the cells genetic response to addition of

nucleotide sugar precursors (Wong et al, 2010).The addition of sodium butyrate is known to alter sialic acid content of the protein and was reported to cause a decrease the expression levels of sialic acid transferase St3gal3 and cytosolic sialdiase neu2 and an increase of lysosomal sialdiase neu1 and membrane sialdiase neu3 (Lee et al, 2014). Unlike the tissue specific glycosylation enzyme, most glycosylation enzymes and transporter expressed in CHO and other industrially important cells are subjected to only relatively small changes of a couple fold. The glycosylation pattern of a product protein is affected by a large number of factors, including the metabolic state of the cell and the supply of nucleotide sugar, the extent that other cellular proteins are also being processed through Golgi apparatus and thus competing for the glycosylation machinery and the abundance level of other components of the secretory pathway. To harness the power of transcriptome analysis and to use changes in the expression profile of genes in glycosylation for steering glycosylation patterns a systems analysis approach using a kinetic model as a predictive tool will be necessary.

3.5 Transcriptomics and Epigenetics for Process Enhancement

During development, the genome is reprogrammed to segregate into regions highly accessible to transcriptional machinery (transcriptionally active) or inaccessible (transcriptionally silent) through chemical modifications of DNA and histone. Such epigenetic alterations are also hallmarks of cancer transformation. Recent findings from the ENCODE consortium studies (Bernstein et al, 2012) have shown the immense impact of epigenetics on gene regulation. The conversion of host cells to a hyper-producer in a short two to three week period during cell line development may also involve genomic reprogramming. While the studies were on human tissues and human cells, the findings are highly relevant transcriptional regulation in recombinant CHO and mouse cells. Here, we highlight a few important findings that may impact biomanufacturing.

The transcript expression level of a gene can be altered through the methylation of CpG (cytidine and guanine dinucleotide)–rich region in its promoter. Methylation of CG sites on the CMV promoter that drives the expression of the product gene was reported to

result in a decrease in productivity in a few cases (Kim et al, 2011; Osterlehner et al, 2011; Yang et al, 2010). In an early application of mouse expression array to cell culture bioprocessing, it was shown that cholesterol auxotrophy in NS0 cells was caused by silencing of the hsd17b7 gene involved in cholesterol synthesis (Seth et al, 2006b). Normal and tumor cells have an inverted methylation patterns. Repetitive DNA regions, usually silenced in normal cells, are hypomethylated in tumors. Overall, there is 20-60% lesser DNA methylation in cancer cells (De Carvalho et al, 2012). How the methylation status of recombinant cell lines used in bioprocessing as compared to normal diploid cells is not known.

3.5.1 Histone Modification

Histones facilitate the packaging of DNA and influence the accessibility to DNA by transcription regulators. Typically, acetylated histone lysines are associated with active chromatin, whereas some methylated histones mark repressed chromatin. Profound changes in histone modification pattern in the pluripotent (and many other) genes upon reprogramming of differentiated cell to iPSC was revealed by Chromatin-immunoprecipitation sequencing (ChIP-Seq) (Mikkelsen et al, 2008). Cancer cells also exhibit a global loss of H4 acetylation and trimethylation (Fraga et al, 2005). Histone modification aberrations in cancer cells result in abnormal gene expression, cell cycle checkpoint instability, and impaired DNA repair (reviewed in (Fullgrabe et al, 2011)). Since these aberrations are also of great concerns in cell biomanufacturing, a better understanding of histone modification in cell line generation and over long-term cultivation is important.

3.5.2 ncRNA

A surprise finding of the ENCODE project is the extensive role played by non-coding RNAs (ncRNA). Although only 2-3% of the genomes encodes for proteins, (i.e., are exons), about 80% of the genome is actively transcribed (Bernstein et al, 2012). While

there are many kinds of non-coding RNA (ncRNA), we will highlight micro RNA (miRNA) and long non-coding RNA (lncRNA).

miRNA are 22 nt long oligonucleotides that associate with the 3'-UTR of its target protein coding genes and inhibit their translation. miRNAs are "global" regulators - one miRNA affects many different transcripts. In cancer, miRNAs can be tumor suppressing (e.g., Let-7 down-regulation increases Ras oncogene expression) or oncogenic (reviewed in (Zhang et al, 2007)). Deep sequencing has been used to identify CHO miRNAs (Hackl et al, 2011; Hammond et al, 2012; Johnson et al, 2011), and more recently, piRNAs (Gerstl et al, 2013). CHO miRNA profiling shows the diversity of expression in different culture phases (exponential or stationary) (Hernandez Bort et al, 2012), growth characteristics (Hackl et al, 2014), conditions such as temperature shift (Barron et al, 2011; Gammell et al, 2007) or nutrient stress (Druz et al, 2011). miRNAs have been reported to delay apoptosis and improve culture performance (Druz et al, 2013; Lim et al, 2006).

lncRNAs (> 200 bp) are usually expressed at lower levels than mRNAs. While a very large fraction of protein coding mRNAs are expressed in all cell types, many lncRNAs are expressed only in a few cell types. They are conserved but seem to evolve faster than protein-coding regions. Recent studies have indicated a wide range of functions for lncRNAs from regulation of gene expression and splicing to imprinting (Derrien et al, 2012). The role of lncRNA has not been reported in cells used in biomanufacturing.

3.6 Transcriptome as a Guide for Cell Engineering

An aim of establishing a hyper-productivity gene set is to devise a highly efficient transcriptome-based cell line development process (Vishwanathan et al, 2013). To achieve a high level of predictability, a very large sample set will be required due to the high dimensionality of the transcriptome data involved. The samples for such studies will likely include cells in the course of transiting from non- or low-producing cells to the final high producing cells. Because of the quantities of cells needed for transcriptome analysis, the cell samples that have been studied in cell line development all had undergone further cell expansion and passed the transient stage of being transformed to become a high producer.

Using microfluidic based RNA-Seq or single-cell transcriptomics (Tang et al, 2009), one will able to discern the transcriptome dynamics during the transformation of host cells to hyper producers.

Transcriptomics may provide insights into cell line screening for specific gene expression levels. Mammals often have different isoforms of the same enzymes that are expressed differently in different tissues to serve various physiological needs. Those isozymes often have different catalytic characteristics and are subjected to different regulations. Cells from different tissues, differentiation and oncogenic transformation states express different isoforms. Different clones of the same cell line, especially those under extensive selection as seen in cell line development for biopharmaceutical production, express different isoforms at different proportions. The role of isozyme variation on the wide range of different behaviors seen in various producing cell lines is not yet studied. They may account for the different metabolic characteristics and the different glycosylation patterns seen in different production lines. A transcriptome-assisted cell line development can, possibly lead to more robust bioprocess development and a better profile prediction of glycosylation or other post-translational modifications. One can envision screening for clones that express fucosylation enzymes at low levels to minimize fucosylated glycans in IgG production for applications relying on antibody dependent cellular cytotoxicity (ADCC) (Niwa et al, 2004; Shields et al, 2002; Shinkawa et al, 2003). In short, the application of transcriptome for biopharmaceutical production will extend from enhancing the productivity to betterment of process robustness as well as product quality and consistency.

With the increased use of transcriptomic tools in cell culture bioprocessing, one will be able to identify more potential targets for cell engineering to enhance process and product characteristics. The advances in genome engineering in the past few years, including zinc finger nuclease (ZFN), transcriptional activator like effector nuclease (TALEN) and CRISPR/Cas9 (reviewed in (Gaj et al, 2013)), will facilitate the transformation of insights from transcriptome analysis to tangible benefits for biopharmaceutical production. ZFN has been applied in CHO cells to create homozygous

deletion of the *dhfr* gene (*dhfr*$^{-/-}$) (Santiago et al, 2008), knockouts of *fut8* (Malphettes et al, 2010), *mgat1* (Sealover et al, 2013), pro-apoptotic genes *bak* and *bax* (Cost et al, 2010) and triple bi-allelic knockout of *dhfr*, *fut8* and *glul* (Liu et al, 2010).

3.7 Conclusion

Transcriptomic studies in conjunction with genome engineering tools have greatly facilitated the development of cell lines for biopharmaceutical production. Microarray and RNA-Seq tools are used increasingly for revealing dynamics and hyper-productivity trait in these cells. Exploration of the epigenetic space will lead to deeper understanding of transcriptome regulation affecting the cells' productivity and stability. With the increasing accessibility and affordability of transcriptomic tools and genomic resources, we can anticipate an increased interest in using such tools for advancement of the biopharmaceutical industry.

Chapter 4: Insights from the Chinese hamster and CHO cell transcriptomes

4.1 Context statement

This chapter is attempt to create a repertoire of transcriptome sequence and expression data for CHO cells. This data will be useful in providing estimates of enzyme expression levels for systems analysis approaches. Systems approaches aid in painting a holistic picture of the cells' state and also help predict their response to different environments. This enables a rational approach to manipulate cellular behavior and enhance their performance for protein production.

Parts of this work are a results of several collaborations and contributions from previous students. Historical RNA sequencing data from Sanger, 454 and Illumina technologies were generated by Katie F. Wlaschin, Cornelia T. Bengea and Nitya M. Jacob, respectively. To the best of my knowledge, earlier versions of the transcriptome assembly were constructed by KFW, followed by NMJ with assistance from Kathryn C. Johnson. Terk Shuen Lee and NMJ annotated the transcriptome assembly used in this section. Even though the latest version of the transcriptome assembly is described in Chapter 7, this assembly was not available to use at the time of writing this manuscript. I collected the data and performed the analysis described in this chapter. I have retained only those sections of work to which I contributed significantly.

4.2 Summary

Systems approaches towards understanding cellular processes have gained popularity in past few years. Especially for CHO cells, a major biotechnology resource, the rich genome and transcriptome information has provided great impetus to such systems-based approaches. Approaches such as modelling the central carbon metabolism or the glycosylation pathway or even building genome-scale models, need estimates for gene expression levels in CHO cell lines to make the models realistic. Our work attempts to

create such a resource in which the range of expression levels for many genes in several different CHO cell lines along with the expression Chinese hamster tissues was collected. This rich information revealed genes that are highly variable in gene expression levels across the CHO cell lines surveyed. Such genes, especially those involved in the energy metabolism and glycosylation are highly relevant for bioprocessing. The comparison of expression levels in CHO cells to that in Chinese hamster tissues, we find that the preference of CHO cells for glycolytic isozymes corresponding to high glycolytic flux. The isozyme expression profile in CHO cells is more similar to brain than liver, and also has a striking similarity to the expression pattern in ovary tissue. A few of the variable genes in glycosylation belonged to the ER glycan processing pathways, Golgi mannosidases and a few sugar transferases. Potential implications of the variability in gene expression is discussed.

4.3 Introduction

Cell lines derived from Chinese hamster ovary (CHO) cells are widely used for the production of recombinant protein therapeutics (Aggarwal, 2014; Kantardjieff & Zhou, 2014). To generate those recombinant protein producing cell lines, transgenes are introduced into a host CHO cell line, such as CHO K1 or CHO DG44, and allowed to integrate into the host chromosome. Subsequent amplification increases the copy number of the transgene and the level of product formation. Finally, clones of cells are isolated for the assessment of their productivity, and the production cell lines are selected after further characterization. For over three decades, the process of developing the production cell lines for a given product has been relatively empirical. To enhance our physiological understanding of cell line development, many have resorted to transcriptome based global survey of producing cell lines and production processes (Kantardjieff et al, 2010a; Korke et al, 2004; Seth et al, 2005; Seth et al, 2007b). EST sequencing projects were also initiated since a decade ago aiming to develop annotated EST sequences and expression microarrays for this economically important species (Becker et al, 2011; Birzele et al, 2010; Jacob et al, 2010; Rupp et al, 2014; Wlaschin et al, 2005).

The rapid advances in DNA sequencing technology in the past decade transformed biomedical research and imparted very notable impact to bioprocess technology. Following the de novo sequencing of the genomes of Chinese hamster and CHO cell lines (Lewis et al, 2013; Xu et al, 2011) and supplementary chromosomal information (Brinkrolf et al, 2013), perhaps dozens of industrial cell lines have been sequenced. The new generation of high throughput DNA sequencing technologies have also made direct sequencing of transcripts, or RNA-seq, readily affordable and applicable to cell line and process investigation (Birzele et al, 2010; Johnson et al, 2013). Over the past decade, through a number of stages of EST sequencing efforts, started with Sanger sequencing, then 454 technology and finally Illumina sequencing, we have accumulated a data set that has a broad range of gene coverage. We have assembled those sequences into contigs and annotated using the homology to mouse and human. With EST and genome sequence data we can also expect increasing applications of proteomic studies on CHO cells since the peptide sequences identified in mass spectrometry can now be validated through sequences translated from EST or genomic data.

The post genomic era in cellular bioprocessing research and development will increasingly adopt a systems approach, integrating genomic, transcriptomic, proteomic and metabolomics data. The recently available Chinese hamster and CHO genomes are likely to play an important role in genome editing for improving protein quality (reviewed in (Baik & Lee, 2014; Steentoft et al, 2014)). Also likely to play an increasing important role are system-wide cellular models, as the complexity of the cellular system will require the aid of mathematical formulation of biochemical and regulatory networks to facilitate our understanding of cellular behavior. In the analysis of the glycosylation pattern of glycoproteins, modeling has been employed to reveal possible alterations under different environmental constrains (Hossler et al, 2007; Jimenez del Val et al, 2011; Krambeck & Betenbaugh, 2005). Using a mathematic model of energy metabolism in mammals it was shown that different combinations of isozymes in glycolysis give rise to very different flux behaviors in glycolysis; with the high flux behavior being the hallmarks of cancer cells (and fast growing cell lines like CHO cells). Glycosylation is a key attribute to the quality

of therapeutic proteins (reviewed in (Gramer, 2014; Jefferis, 2009)). The metabolic behavior of cells in fed batch cultures has been reported to greatly influence the process outcome (Burleigh et al, 2011; Chee Furng Wong et al, 2005; Ha & Lee, 2014; Liu et al, 2014; McAtee et al, 2014; Nyberg et al, 1999; Seo et al, 2014). A systems biotechnology approach employing models that describe various cellular dynamics, will greatly enhance our ability to modulate the glycosylation profile of therapeutic proteins or to control cellular metabolism for increased process robustness.

In developing a mechanistic model and in undertaking a systems approach, one frequently encountered hurdle is the estimation of the abundance level of the involved proteins (enzymes, transporters, binding proteins etc.). When multiple isoforms are involved in a reaction step, it is required to identify the isozymes expressed in the cell type to be studied, and their abundance levels. Given that the makeup of isozymes in a pathway and the level of enzymes in a network affects the behavior of biochemical systems profoundly, an evidence (or experimental data) based estimate of the expression levels of those key proteins will help in establishing systems tools for CHO cells.

Herein we compiled the transcript expression data from RNAseq and DNA microarray studies on the genes in a number of pathways in CHO cells lines to allow for an order-of-magnitude estimate of their corresponding protein levels. We highlighted their characteristic isoform expression by comparing their expression levels to liver and brain tissues. The compiled data shows a high degree of variability of gene expression levels among different cell lines. Such variability may contribute to the phenotypic variability among different cell lines or cell clones. Another possible source of variability in cellular response in culture is mutation, especially in the gene coding region. Cells in culture may accumulate mutations in the population. CHO cells have been cultured extensively since their isolation decades ago (Puck et al, 1958; Tijo & Puck, 1958). The RNAseq data covered a large number of genes at the transcript level to very high depth. Sequence variants, even at low levels, can thus be detected. We thus also analyzed the frequency of sequence variants among the seven cell lines for which RNA-seq data are available.

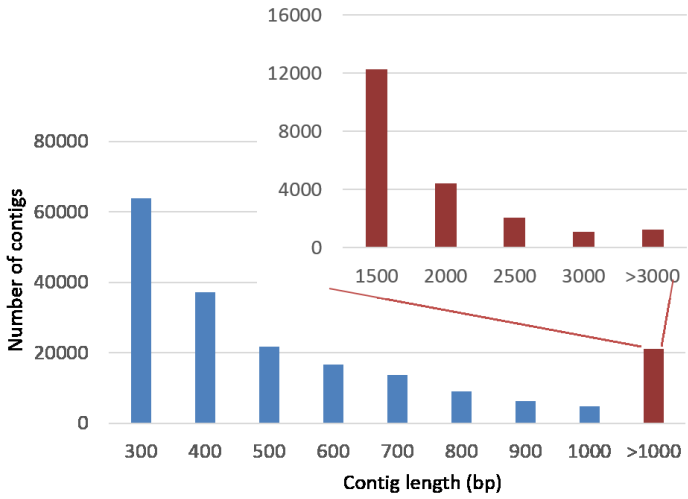
4.4 Materials and methods

4.4.1 Source of Transcripts and EST Sequencing

Complementary DNA (cDNA) libraries from Chinese hamster tissues and different CHO cell lines were constructed for sequencing using 454 and Illumina sequencing technologies. The average read length of 454 sequencing was 210 bp and for Illumina sequencing was 90 bp. A total of 40 Gbp of sequencing data was used in this study (Appendix Table 0-3).

4.4.2 *De novo* Assembly of CHO Transcriptome

The sequencing data was supplemented with Sanger sequencing reads reported previously (Jacob et al, 2010; Wlaschin et al, 2005). The redundant reads were removed by mapping to the assembly of the Sanger reads. Using a combination of assemblers like Velvet (Zerbino & Birney, 2008) and MIRA (Chevreux et al, 1999) to assemble the diverse nature of sequencing reads, Illumina and 454, respectively. We therefore used Phrap, a long read assembler, to integrate these contigs together to obtain longer sequences of transcripts. The final CHO EST collection consists of 194,450 sequences, with a median length close to 600 bp (Figure 0-7).



No. of Contigs	194,450
Total Transcriptome Length (bp)	108,119,200
Median Contig Length (bp)	387
Average Contig Length (bp)	556
N50 Contig Length (bp)	671
Maximum Contig Length (bp)	15,140

Figure 0-7: Chinese hamster and CHO transcriptome assembly. Final transcriptome assembly statistics-distribution of length of contigs. Inset shows the contig length distribution for the contigs greater than 1000 bp. Table shows more detailed statistics.

4.4.3 Annotation of the CHO Transcriptome

Functional annotation of the assembled CHO ESTs was based on the identification of orthologs in closely related species. We employed a tier-based annotation strategy and used several public databases for this process, including ENSEMBL (Mouse, Rat, Human) (www.ensembl.org), FANTOM3 (Carninci et al, 2005), NONCODE and GenBank. Repeat-masked ESTs were first aligned to ENSEMBL Mouse, FANTOM3, ENSEMBL Rat, ENSEMBL Human and NONCODE (Figure 0-8) using NCBI BLAST. Those that did not align to any sequence in these databases were searched against the GenBank and RefSeq nucleotide repositories. The details of the pipeline are described in Chapter 7.1.

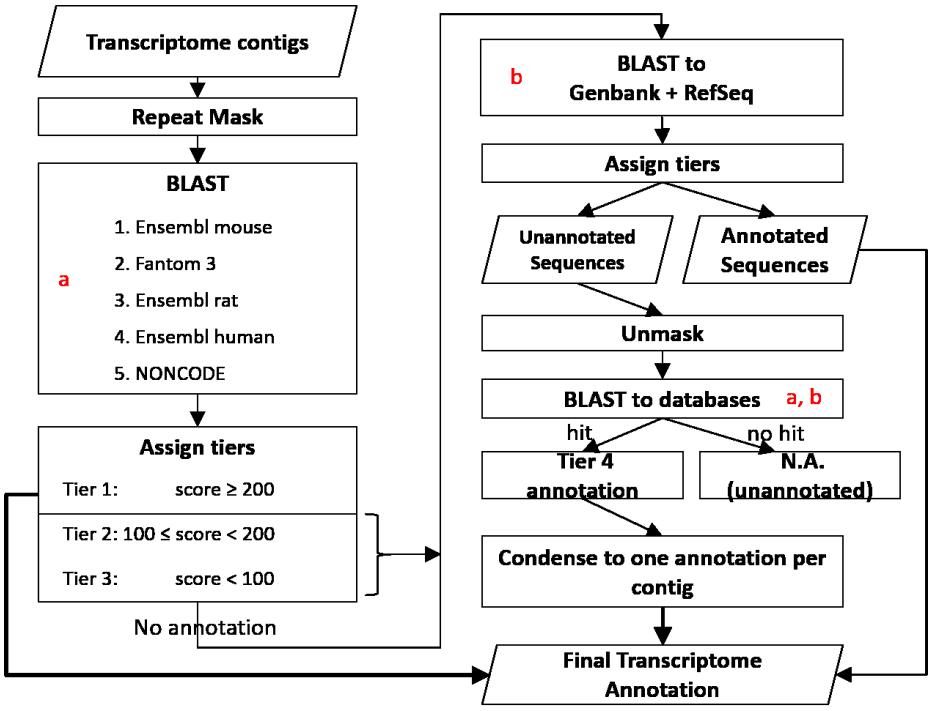


Figure 0-8: Transcriptome annotation schema showing homology-based annotation strategy. The annotation was divided into tiers based on the strength of the annotation.

Finally, any transcripts which could not be annotated via this pipeline were searched for the presence of open reading frames and protein motifs from the PFAM database.

4.4.4 RNA-Seq based quantification of gene expression

The transcriptome sequencing data from six cell lines and two tissues were mapped to the annotated transcriptome using Bowtie software (Langmead et al, 2009) allowing for 3 mismatches per 90 bp read. The number of reads mapping to a contig were summed up to generate counts for each contig. The counts were then normalized across nine libraries to adjust the expression values for variability in sequencing depth using the upper quantile normalization procedure (Bullard et al, 2010). The upper quantile normalized values were further adjusted for variability in contig lengths. For each Ensembl-annotated gene, the

contig with the maximum number of reads mapped was considered representative of that gene for further analysis.

4.4.5 Microarray

Based on the assembled and annotated transcriptome, a microarray was constructed using the NimbleGen platform.(Jacob, 2011). Liver, brain and ovary tissues and parental CHO cell lines K1, DXB11 and DG44 along with 6 recombinant cell lines (four derived from DXB11, two from DG44). Three of these produced IgG, two produced a TNF-alpha Fc fusion protein derived from DXB11 and one remaining expressed only DHFR gene. Three of these lines were additionally treated with methotrexate, one was treated with 2 mM sodium butyrate for 24 hr prior to sourcing the RNA. All RNA samples were taken from cells at exponential growth stage, except one was in late stage. A total of 14 samples from cell lines in addition to three from Chinese hamster tissues were used for RNA extraction (Appendix Table 0-3). RNA was extracted from samples using the RNeasy Mini kit (Qiagen) using standard manufacturer-recommended protocols. cDNA was synthesized from the samples, labeled, and hybridized to the custom microarray. Data obtained was linearly normalized to a mean expression value of 500.

4.4.6 Hierarchical clustering

The linear normalized data was clustered using UPGMA (un-weighted average) clustering method in Spotfire DecisionSite (TIBCO, Somerville, MA). The Euclidean distance metric was used to estimate the inter-cluster distance.

4.5 Results

4.5.1 The CHO Transcriptome

We have performed RNAseq on the mRNA samples from liver and brain tissues of Chinese hamster and nine CHO cell lines, assembled them into contigs (see Materials and Methods, Figure 0-7) and subsequently annotated them based on homology to mouse and human sequences (Figure 0-8). Approximately half of the EST contigs were annotated

using the two mouse databases (ENSEMBL Mouse, FANTOM3) with a stringent E-value cutoff of 1e-25. Furthermore, another 20% of the total sequences could be confidently annotated against the other databases. When the E-value cut off was relaxed to 1e-4, a further 27% of the ESTs in our collection could be assigned annotation. Overall, 63% of the sequences were assigned tier 1 annotation, of which ~70% have ENSEMBL mouse annotation (Figure 0-9A).

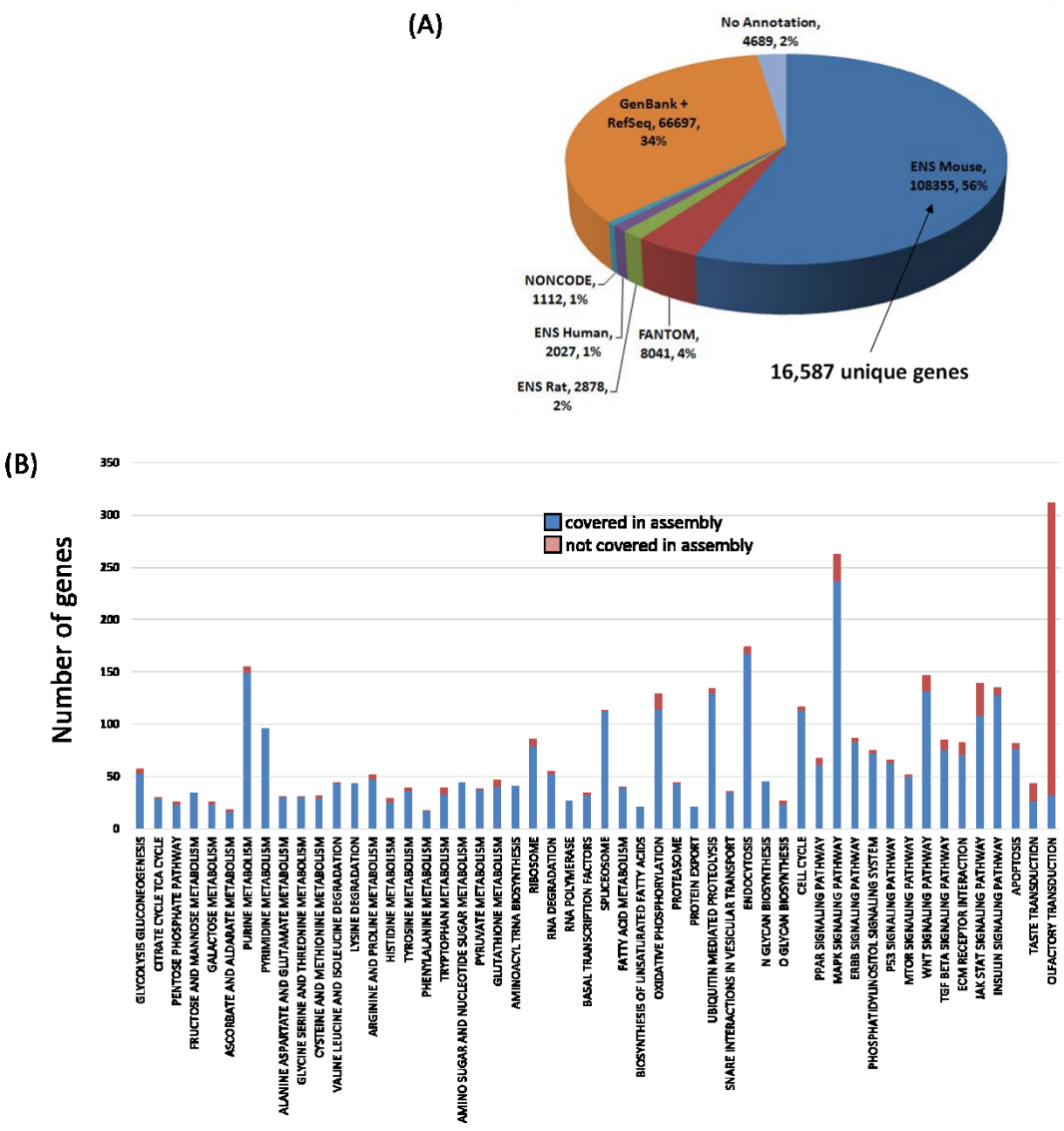


Figure 0-9: (A) Distribution of contig annotations among the reference databases. (B) Coverage of genes in major pathways.

Comparing the EST collection to the 22,036 Chinese hamster/CHO sequences present in RefSeq, 77% of the RefSeq sequences are represented with an E-value < 1e-22. The sequences are covered by 77,244 out of 194,450 EST contigs in our collection. The coverage of major pathways was assessed. The list of genes in each pathways was obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Most of the pathways that are relevant in the context of bioprocessing, such as N-glycosylation, energy metabolism glycolysis, tricarboxylic acid cycle or apoptosis, are almost completely covered. A few pathways involved in signaling, MAPK pathway, phosphatidylinositol pathway have a larger number of genes not yet assembled or annotated. Other pathways with lower coverage are from the taste transduction pathway and olfactory transduction pathways (Figure 0-9B).

4.5.2 Comparative Analysis Of Gene Expression Between Chinese Hamster Tissue and CHO Cell Lines

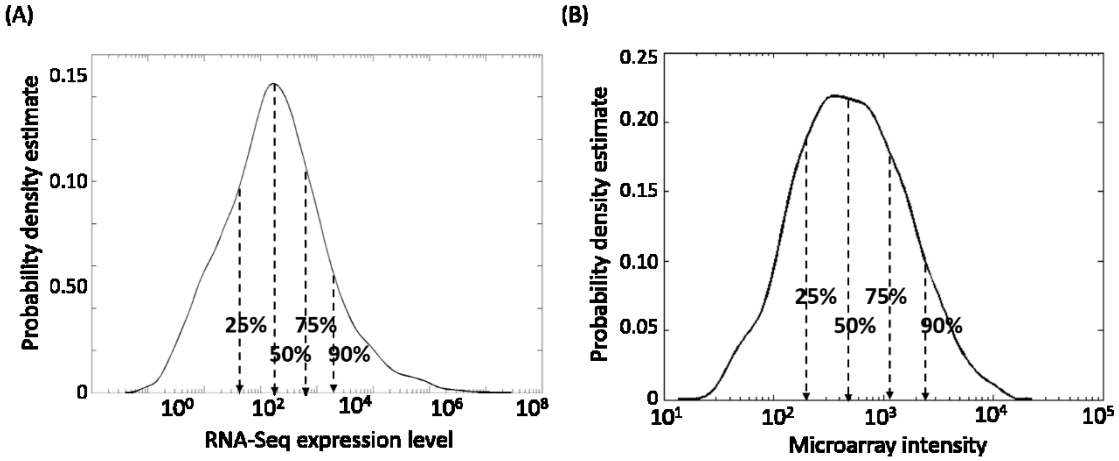


Figure 0-10: Distribution of expression levels in (A) RNA-seq and (B) Microarray. Arrows indicate 25%, 50%, 75% and 90% of the data from the cumulative distribution.

The annotated transcriptome sequences were then used as a reference for quantification of the expression levels of genes in Chinese hamster liver, brain and that of

the six cell lines using RNA-Seq reads. Using the microarray constructed with the assembled contigs, the expression levels of genes in major pathways for fourteen samples consisting of nine cell lines under different conditions was assessed (Appendix Table 0-3). The use of these rather diverse samples was intended to give a representative view of the range of transcript level for genes that are relevant for bioprocessing. To give a context of relative expression levels the distributions of RPKM value and microarray intensity value are shown in Figure 0-10. The values for each quartile, and top 10% are also marked in the graph. In the section below we will first compare the expression level of enzymes in energy metabolism in CHO cells and tissues. We will next describe the divergence of transcript levels among CHO cells in other pathways bioprocess significance.

4.5.2.1 Glycolysis Enzymes

The transcript level of enzymes in glycolysis, in liver and brain tissues as well as the average of all cell lines as determined from RNA-seq are shown in Figure 0-11A. In this study RNA-seq was not performed on RNA sample of Chinese hamster ovary. However, we performed microarray transcriptome analysis on ovary tissue. The transcript expression levels of glycolysis genes of ovary tissue along with three CHO cell lines are shown in Figure 0-12. Several enzymes in glycolysis, hexokinase (HK), phosphofructokinase (PFK1), 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase (PFKFB), enolase (Eno), pyruvate kinase (PK) and lactate dehydrogenase (LDH), have isoforms that have different kinetic behaviors and are subject to different allosteric regulations. The combination of isozymes expressed in different tissues gives each tissue its metabolic characteristics.

The dynamic range of transcript levels of glycolysis genes spans over 30 fold in brain, liver and CHO cells. As is well known GAPDH is the highest expressed in all cells and tissues. In fact, it is among the top 1 % highest transcripts in every tissue and cell line surveyed. In liver adolase (aldo) and fructose bisphosphate phosphatase (FBP) are also expressed at very high levels, approaching that of GAPDH. In all other samples, a group of enzymes are expressed at levels of 1/5 to 1/10 of GAPDH, including enolase, pyruvate

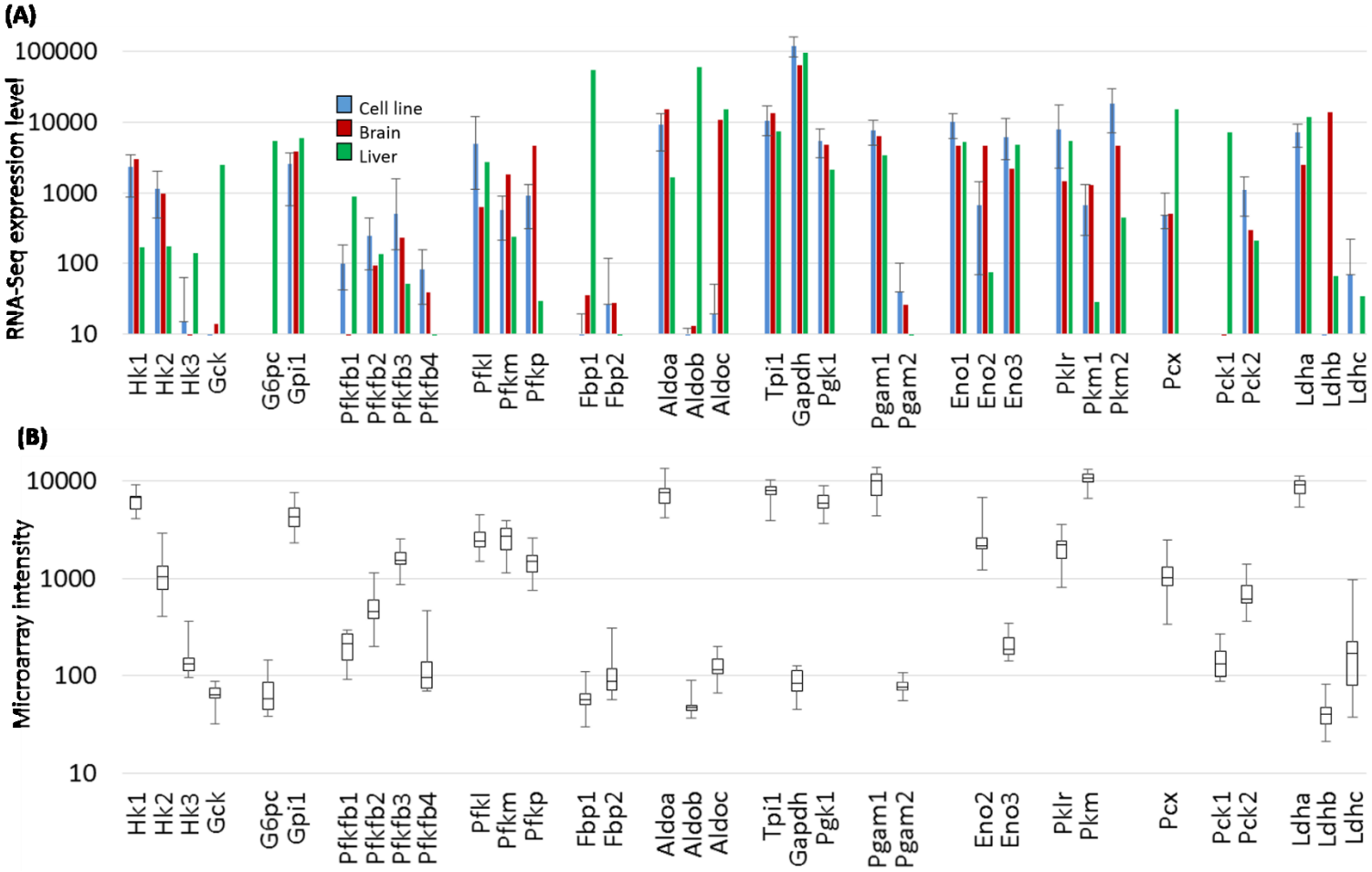


Figure 0-11: Glycolysis gene expression levels from (A) RNA-Seq expression levels of cell line average, liver and brain. The bar indicates the range of expression in cell line. (B) Box-plot of microarray expression data for 14 cell line samples.

kinase pfk, pgk, pgm, tpi and ldh. Glucose transporter and glucose-phosphate isomerase (gpi) levels in brain and cell lines are lower than all other enzymes. In liver, these two genes are expressed at higher levels. Pfkfb, being an enzyme playing primarily a regulator role, has a wide range of expression levels in different samples.

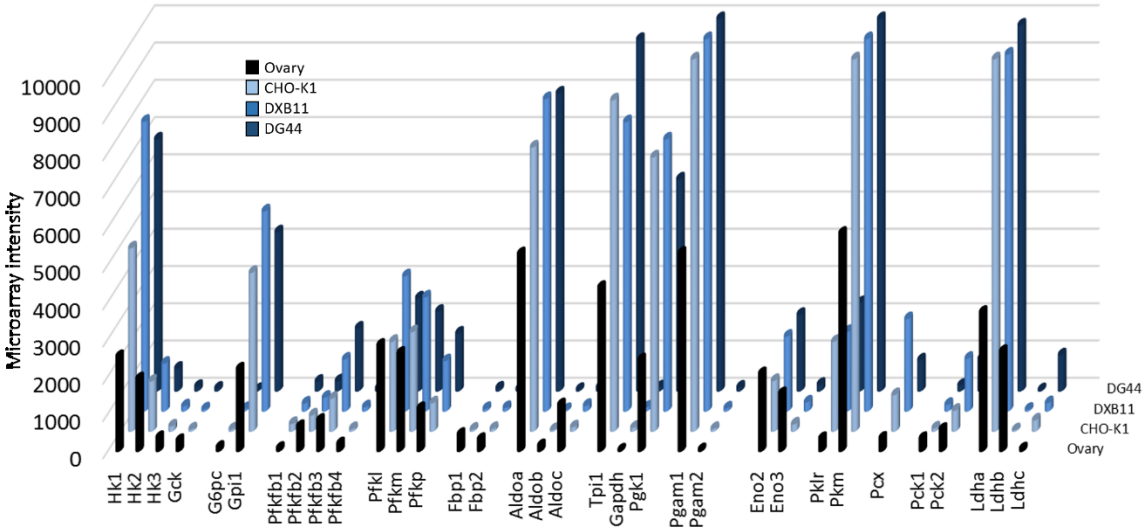


Figure 0-12: Expression levels of glycolysis enzymes in parental CHO cell lines compared to ovary expression level measured by microarray.

As is expected the pattern of isozyme expression in CHO cells is very distinct for tissues. Four isoforms involved in gluconeogenesis, glucokinase (gck), glucose 6-phosphatase (g6pc), phophoenolpyruvate carboxykinase (PCK) and pyruvate carboxylase (PCX) are expressed at high levels only in liver. Liver also expresses liver isoform of pyruvate kinase (pkl) and pfkfb1 almost exclusively (the other isoforms are expressed at least ten times lower level), however these liver isoforms are also expressed in brain and CHO cell as minor isoforms, Pfkfb1 is a bifunctoinal enzyme that catalyzes the formation of fructose 2,6-bisphophate from fructose 6-phosphate (the kinase activity) and the reverse reaction of hydrolysis of f26p to f6p (phosphatase) has a higher kinase to phosphatase activity. The high expression level of pfkfb1 in liver gives it the capability of rapidly changing glycolysis flux, even to revert it to gluconeogenesis.

Ldhb is uniquely expressed in brain. Brain expresses the platelet form of phosphofructokinase (pfkp), while liver and CHO expressed predominantly the muscle and

liver forms (pfkm and pfkl). Pfkl and pfkm are subjected to activation by fructose-1,6-bisphosphate, however, this interaction is almost absent for pfkp. CHO cells and liver have high levels of pfkm/pfkl. They thus have a high glycolytic rate as a result of F16P activation. Hexokinase isozymes, hk1 and hk2, are highly expressed in CHO cell lines and brain. Both hk1 and hk2 have a low Km and have the ability to physically bind to the outer membrane of mitochondria. This association with mitochondria confers hk1 and h2 direct access to ATP generated by mitochondria and provide driving force for high rate of glycolysis as seen in both brain and CHO cells.

Overall the isoenzyme pattern of CHO cells is closer to brain than to liver. Notable differences between CHO cell and brain are ldh (ldha isoform in CHO cell while ldhb in brain) and pfk (pfkl in CHO cell and pfkp/pfkm in brain). Interestingly from the microarray data the isozyme expression pattern between CHO cell line and ovary are rather similar except that ovary also expresses Eno3 (Figure 0-12). Pkm1 and pkm2 are alternatively spliced forms differing in only one exon, originating from the pkm gene, with a difference in 23 amino acids. Pkm2 is activated by fructose-1,6-bisphosphate, but not pkm1. Pkm1 is expressed in majority of adult tissues, while pkm2 is restricted to embryonic tissues and transformed cells. CHO cells express pkm2 as the dominant isoform. The microarray data show that ovary express pkm also, however, the microarray probe design does not allow for differentiation of pkm1 and pkm2.

4.5.3 Variability of transcriptome profiles among cell lines

4.5.3.1 Glycolysis genes

To examine the variability of the expression level of genes of glycolysis in different cell lines, a box plot of the transcript level obtained from microarray on all fourteen cell lines are shown in Figure 0-11B. All the cell lines surveyed express the same dominant isoforms of the glycolysis isozymes. However the transcript level of individual enzymes and the ratio of different isoforms of a given enzyme vary rather widely among cell lines. This is seen in enzymes which play major regulatory roles, such as pfkfb, pk, as well as in other enzymes. CHO cell lines often behave differently in their metabolic profile in terms

of their consumption of nutrients and production of metabolites. The variability of enzymes involved in glycolysis may play a role in the metabolic diversity.

4.5.3.2 N-Glycosylation genes

The glycosylation pattern of the therapeutic protein products produced by CHO cells is a defining parameter of product quality. The expression of glycosylation enzymes can impact on the glycosylation profile of the product proteins. We examined the transcript levels of about fifty N-glycosylation enzymes expressed in CHO cells as assayed with microarrays in fourteen samples (Figure 0-14). The pathway starts from adding the first two GlcNAc residues to the dolichol phosphate and growing by transferring mannose to form the initial oligosaccharide precursor ($Man_5GlcNAc_2$) in the cytosolic side of the ER membrane. This is followed by flipping the oligosaccharide into the ER lumen, and subsequent growth to 9-mannose ($Man_9Glc_3GlcNAc_2$) oligosaccharide precursor. The Alg family of enzymes play key roles in the synthesis of the oligosaccharide backbone. The glycan precursor is transferred to nascent protein molecule and subsequently to Golgi apparatus for further glycoprocessing as N-linked glycan (Figure 0-13). The microarray intensity data indicates that most of the genes involved in the glycan precursor formation (Alg family and Dpm genes) are expressed at moderate levels (ranging from 200-2000 with a median intensity of 500 of all transcripts). The transcript levels are variable among cell lines, but not particularly varying over wide ranges compared to other genes. The Oligosaccharyl transferase (OST) complex consisting of stt3a, stt3b subunits, and associated protein tusc3 are involved in the transfer of the oligosaccharide to the nascent protein molecule. The enzymes involved in the glycan translocation to nascent proteins are all expressed at higher levels of about five fold higher than the Alg family proteins. Many of those enzymes also have a wide range of expression level (Figure 0-14).

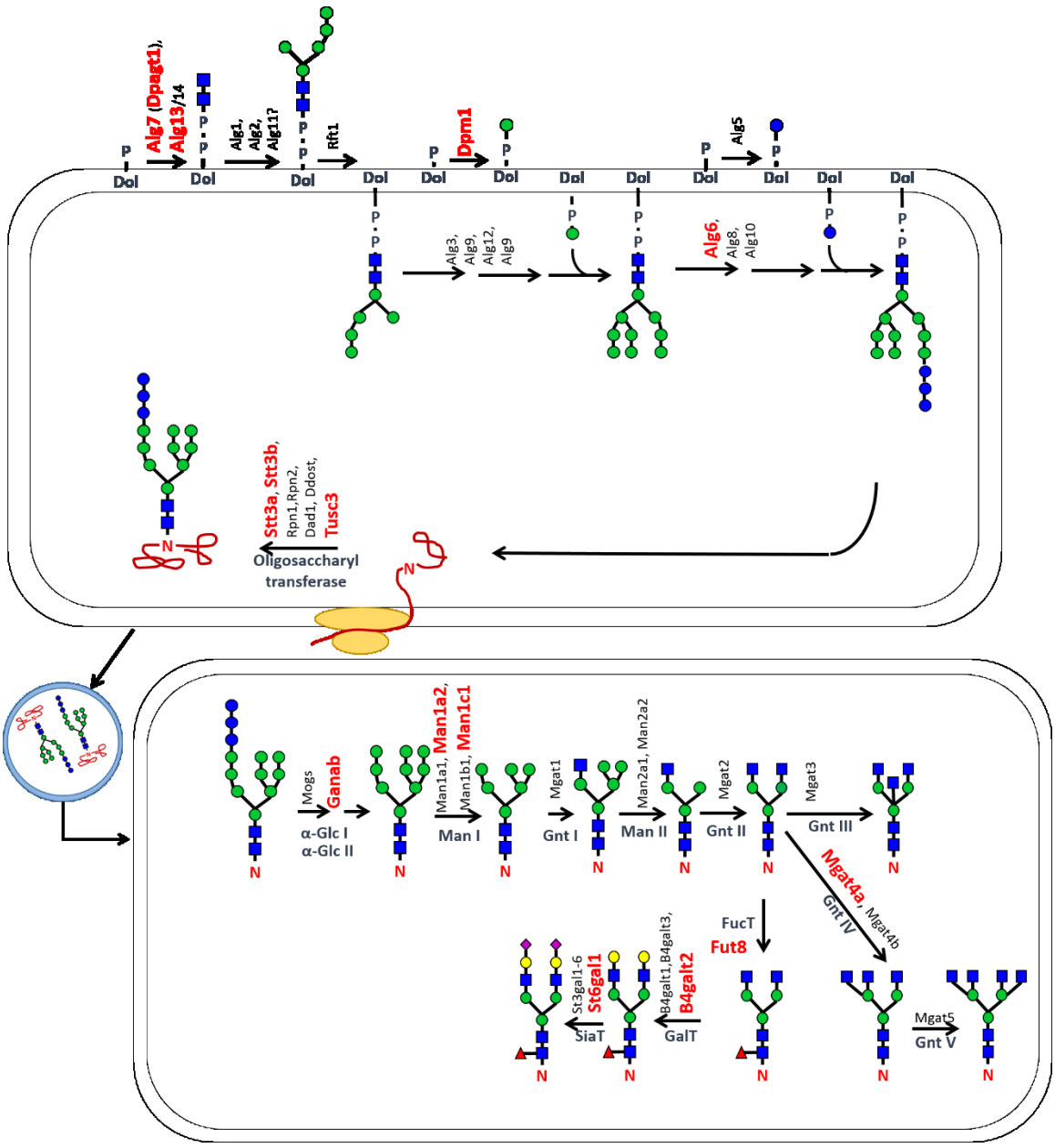


Figure 0-13: Protein N-glycosylation pathway. The genes that are variable in expression level are indicated in red font on the pathway.

Subsequently the glucose and mannose residues in the oligosaccharide are trimmed. The glucose trimming enzyme, Ganab, is also expressed at a rather high level, but not the mannose trimming α-(1,2)-mannosidases Man1a1, Man1a2, Man1b1 and Man1c1 (Figure 0-14). Members of the mannosidase family have been reported to be among the

hyperproductivity gene set in the study of high producing cell lines (Charaniya et al, 2009). Some of those enzymes (Man1a2, Man1b1) have very wide range of transcript levels among cell lines, spreading over 10 fold.

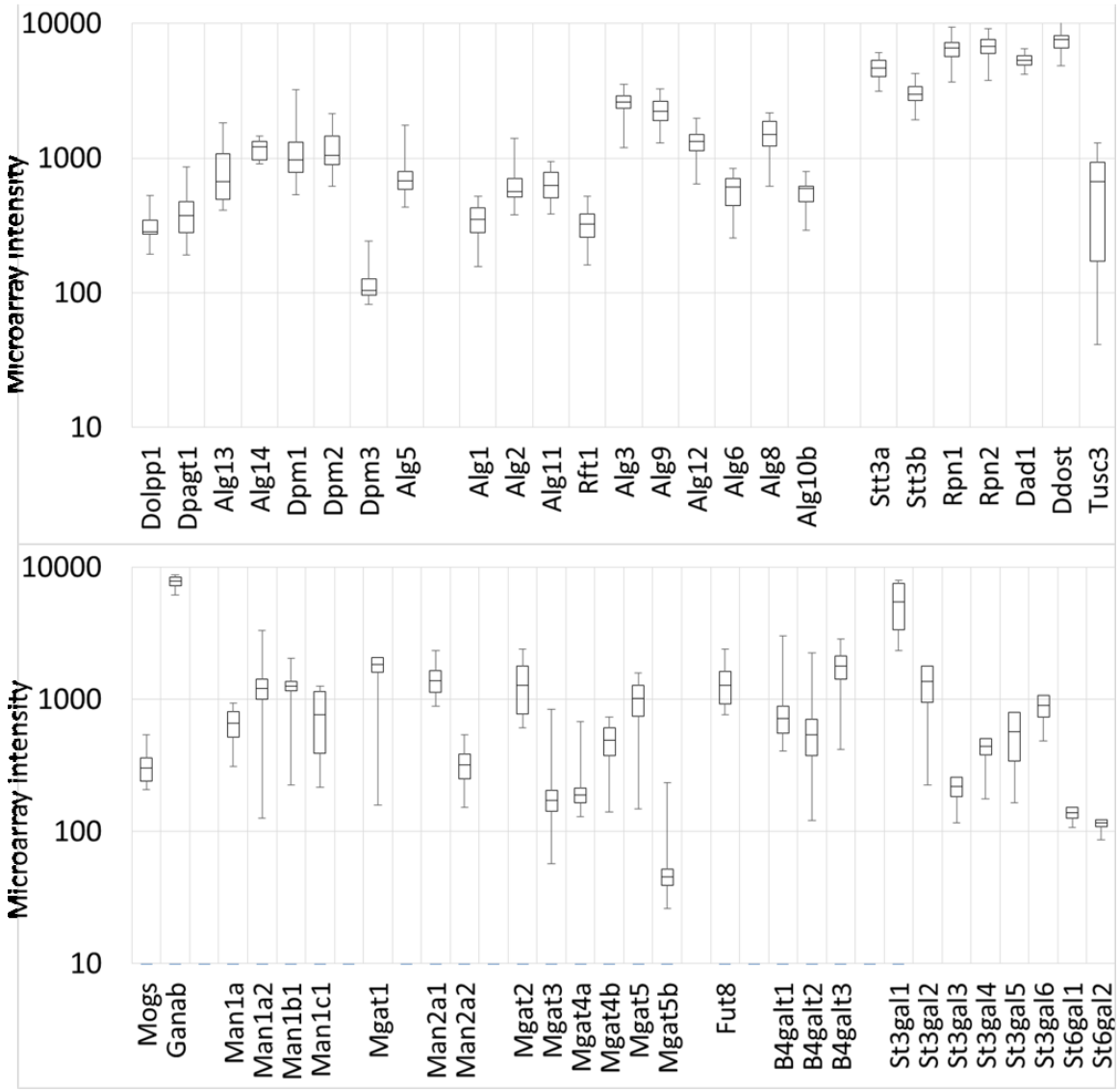


Figure 0-14: Expression levels of glycosylation enzymes in 14 cell lines represented as a box plot. Top panel are ER localized enzymes, and bottom panel has the Golgi localized enzymes.

Most of the Golgi glycosyltransferases in CHO cells are expressed at levels comparable to those in brain or liver with three exceptions of beta-1,4-mannosyl-4-beta-

N-acetylglucosaminyltransferase (Mgat3), and two sialyltransferases (St6gal1 and St6gal2). Mgat3 adds the bisecting GlcNAc to the oligosaccharide and showed extremely low expression in all the CHO cells (Figure 0-14). The presence bisecting GlcNAc enhances antibody dependent cellular cytotoxicity (ADCC) of recombinant immunoglobulin G and is potentially a desirable feature for some therapeutic antibodies.

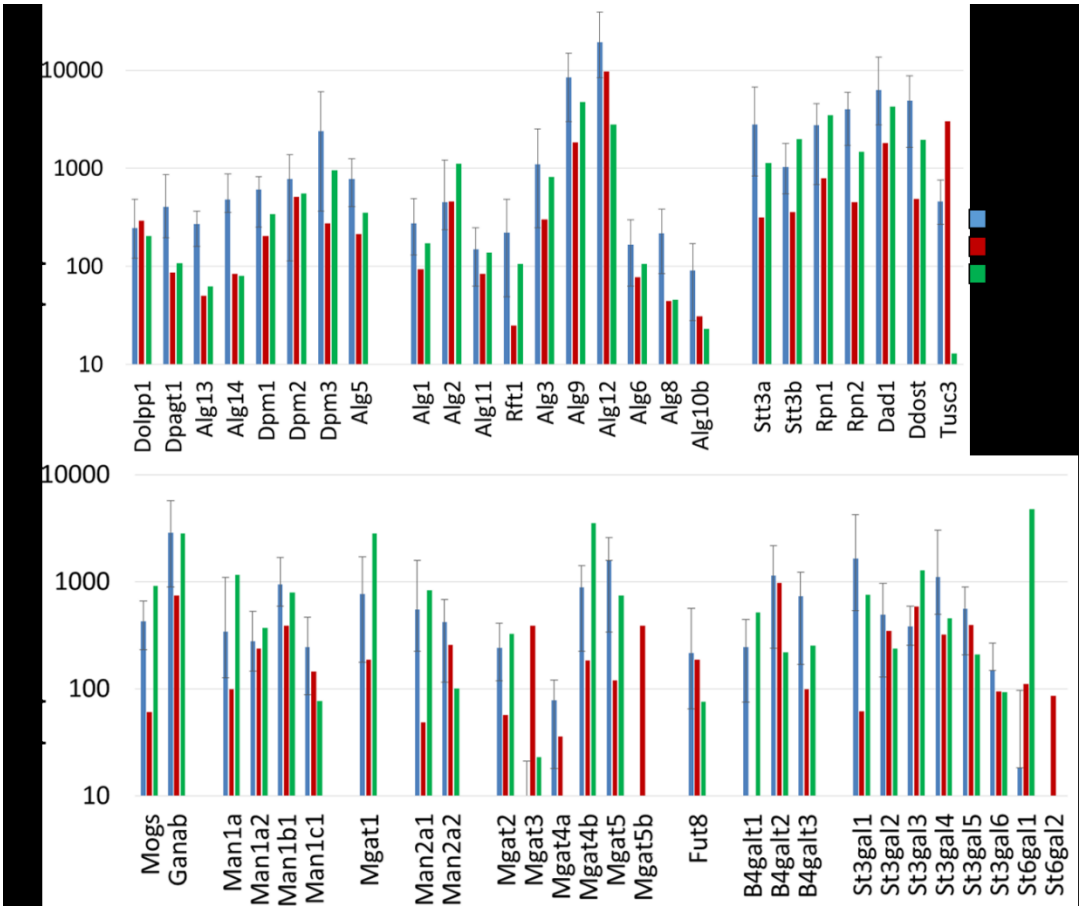


Figure 0-15: RNA-Seq expression levels of glycosylation genes for cell lines, and brain and liver tissues. The bar on the cell line (average of 6 samples) represents the range of expression.

The α-(2,6)-sialyltransferase genes St6gal1 and St6gal2 show almost no expression in CHO cells, but St6gal1 is very at a very high level in liver (Figure 0-15). In contrast, the α-(2,3)-sialyltransferase genes (St3gal family) are highly expressed in CHO, liver and brain tissues. This transcript pattern is reflected in the low 2,6-sialyl and high 2,3-sialyl glycan content in recombinant proteins produced in CHO cells. St3gal1 is the dominant α-(2,3)-

sialyltransferase (Figure 0-15), however, its expression level is highly variable, spanning over five fold, from a moderately high level (intensity 2000) to very high level (intensity almost 12000).

The expression level of Fut8 in CHO cells, in the range of moderate level of 500-1000, is higher than that in liver Figure 0-15. Lower expression of Fut8 has been shown to correlate with increased recombinant IgG therapeutic performance.

4.5.3.3 Heritability of transcriptome signatures

The transcript profiles of all cell lines and tissues obtained from RNA-seq and from microarrays were subjected to hierarchical clustering using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm. As can be seen in Figure 0-16, cell lines from the same lineage or derived from the same parental source were clustered together, using both microarray data and the RNAseq data. With microarray data cell lines which originated from the same parental lineage also cluster together; those derived from DG44 formed a cluster in spite of their having been subjected to highly stressed conditions of transgene amplification as in the cell lines. One DG44-derived recombinant cell line did not cluster together along with its group. One of the DXB11-derived cell lines that was treated with methotrexate, also did not cluster with the DXB11 cluster. At a higher level, the cell lines are clustered together with ovary tissue, forming a distinct group from brain and liver.

In a separate study in our lab, the SNVs that have sufficient depth of coverage in all cell line libraries were also analyzed for tracing cell lineage. The resulting tree grouped cell lines derived from the same parent together. Cell lines derived from the same parent retain the common variant markers in their population. It is interesting that both the transcriptomes, and the profile of transcript expression levels of all genes, and nucleotide variants allowed for tracing of lineage of cell lines.

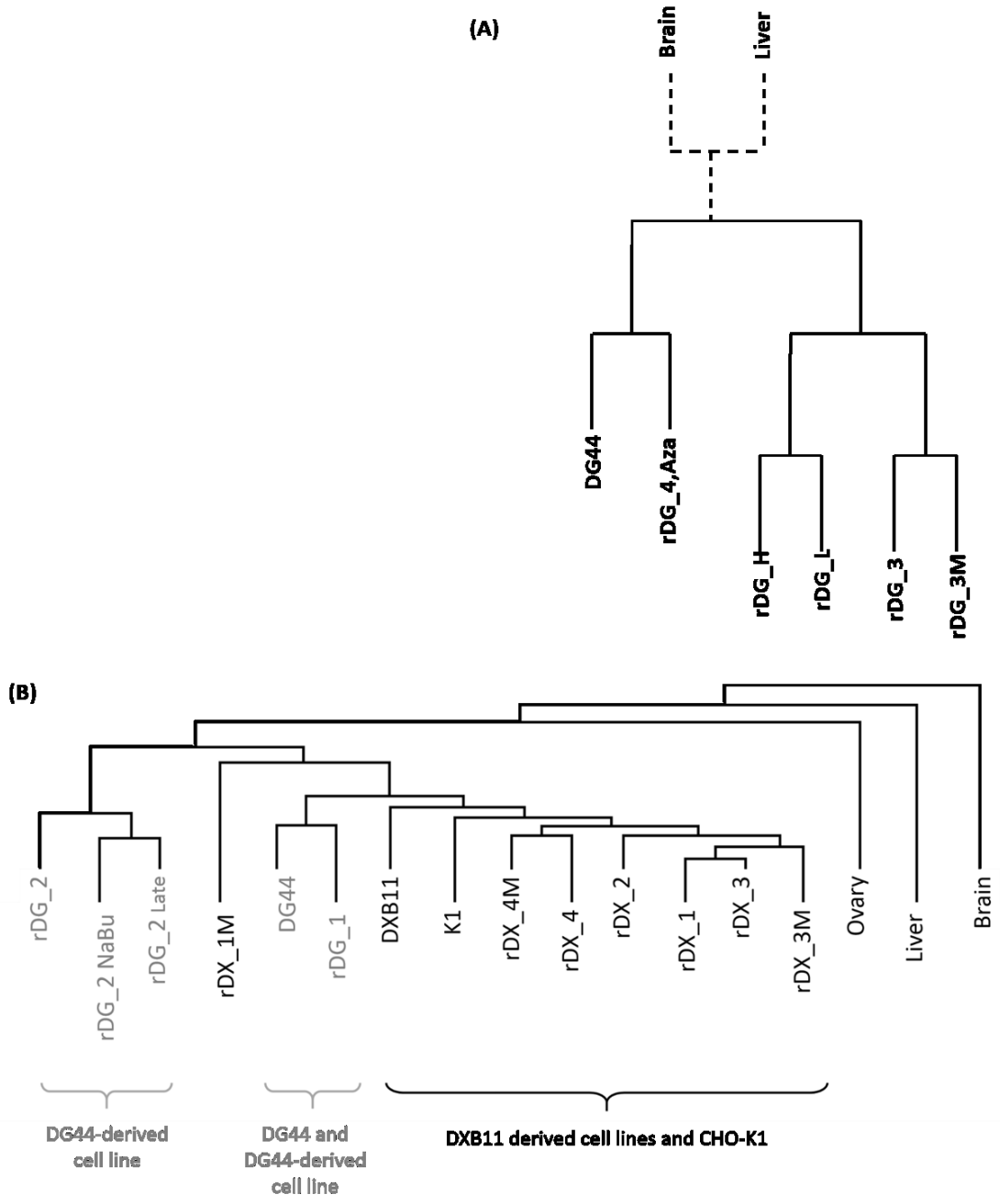


Figure 0-16: Hierarchical clustering of expression data from (A) RNA-seq samples and (B) microarray samples.

4.6 Discussion

Since the beginning of transcriptomic exploration in CHO cells a decade ago an increasingly large size of transcriptomic data has been accumulating. In the past few years the pace of accumulation has greatly accelerated due to the rapidly decreasing cost of high throughput DNA sequencing. The RNA-seq data that were acquired in our previous studies were first assembled and annotated before the transcript levels for different genes or contigs could be obtained. The annotated ESTs were then used to construct a transcription gene expression microarray for this study. The two transcriptome surveying methods employed are complementary. RNA-seq allows for higher accuracy of transcript identification by mapping the reads to the entire transcript. We thus employed RNA-seq data to identify isozymes, even those which are alternatively spliced, expressed in CHO cells. We then take advantage of the broad gene coverage of microarray to obtain a comprehensive gene expression data on a diverse cell lines economically.

The global nature of transcriptome has demanded a more complex and systematic means of transforming the data into physiological information. Increasingly the understanding of cell's behavior is hinged upon the development of models for systematic analysis. Notable examples are the metabolism and glycosylation pattern in the production of recombinant proteins. The stability behavior of energy metabolism and the flux distribution in glycan biosynthesis cannot be prescribed without resorting to a systems approach. Important to establishing a descriptive model, is the estimation of the enzyme or other protein levels for model analysis. Many kinetic constants can be measured, or obtained from literature, or estimated from reactions of similar nature. However, protein levels for a biochemical reaction system often cannot be easily assessed. With the transcriptome data compiled from RNAseq and from microarrays we identified the isozymes and their range of expression in pathways. These values can be used as a first order estimate of the relative protein levels for systems analysis. In the above section we described glycolysis and glycosylation pathways. Data of other major pathways important to CHO cell physiology are given as supplementary data.

The composition of isozymes in glycolysis has a profound effect on the metabolism of tissues and cells. It was shown recently that, depending on make-up of isozymes, cells may reside in a high flux or a low flux state at the given glucose concentration,(Mulukutla, 2012; Mulukutla et al, 2012). The glycolytic isozyme profile of CHO is closer to brain than to liver. It is not surprising that pkm2 is the dominating isoform of PK. What is unexpected is the very high level of the isoform and the wide margin of expression levels among different cell lines. Overall the variability of gene expression level in glycolysis is high. Pfkfb, that plays an important role in regulating glycolysis flux, also shows variable expression among cell lines. Different isozymes of pfkfb exert different stimulatory effect on glycolysis flux, while different levels of the same isoforms of pfkfb affect the speed at which metabolism responds to changing environment especially glucose levels. The variability in the expression levels of pkm2 and pfkfb may contribute to the diversity in metabolism among cell lines.

Most glycosylation enzymes are expressed at levels comparable to native tissue expression level. However, the expression levels of α-2,6-sialyltransferases St6gal1 and St6gal2, and bisecting GlcNAc transferase Mgat3 are at lower levels than native tissue. Interestingly, the region upstream of the St6gal1 gene has four CpG islands. On interrogation of the methylation status of the CpG islands, and found that most of those four CpG islands in CHO cells are methylated (unpublished work). In contrast, those in liver are not methylated. The methylation status is thus consistent with the transcript expression profile. It is thus possible that St6gal1 in CHO cells can be de-silenced to produce product proteins with sialic acid more similar to human form. Elimination of expression of Mgat5 can reduce the dimensions of the Golgi apparatus (Dong et al, 2014), suggesting that glycosyltransferases may have functions other than participating in the building of the glycan.

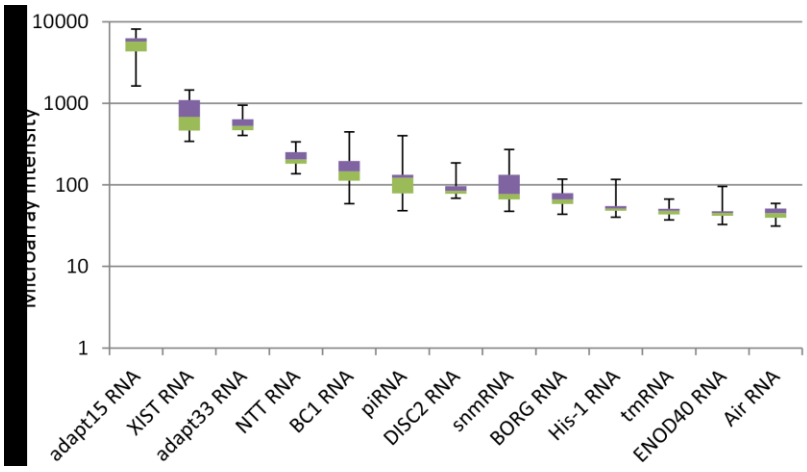


Figure 0-17: Long noncoding RNA expression levels among CHO cell lines.

A number of non-coding RNA were among the transcripts profiled in RNA-seq and microarray assays. The roles of long ncRNA in regulating cellular physiology are only beginning to be revealed even in the better studied human (Derrien et al, 2012) and mouse. Only a few of transcript are annotated as ncRNA in our Chinese hamster transcriptome. Non coding RNAs adapt15 and adapt 33, that were earlier shown to be associated with oxidative stress response in hamster fibroblasts (Crawford et al, 1996a; Crawford et al, 1996b; Wang et al, 1996; Wang et al, 2003), are found to be highly expressed in all CHO cells (Figure 0-17). Xist, a non-coding RNA that regulates X-chromosome inactivation for dosage compensation in mammalian females, is also found to be expressed significantly in CHO cells. Ntt (non-coding transcript in T cells), BC1, DISC2 (disrupted in schizophrenia 2), BORG (BMP/OP-responsive gene), His-1 and Air (antisense Igf2r) are a few other non-coding RNAs expressed in CHO cells. Apart from these, there are a lot of unannotated ncRNAs found to be expressed in CHO cells. What role those ncRNA may play in CHO cells is an area open for future investigation.

The RNA Seq sequenced the transcripts to varying depth depending on the abundance level of each gene. While in genome sequencing the sequencing depth ranges from a few fold in depth for re-sequencing to 60-100 fold for de novo sequencing, the RNAseq performed in our study reach thousands to tens of thousands fold in depth for

abundant genes. The RNAseq data can thus be exploited to examine the sequence diversity to a greater detail. The recombinant CHO cell lines have undergone selection, even transgene amplification and single cell cloning that at some point of their derivation refined each line's clonality. When a population of cells undergo single cell cloning, the mutations accumulated in the population is potentially drastically trimmed in size. The cell lines we examined in RNAseq involves both parental cell lines as well as recombinant cell lines that were once cloned.

In conclusion, an order of magnitude estimate of the gene expression levels in CHO cells will be a great aid in systems biology approaches. The systems biology experiments give a holistic understanding of the system. This will enable more rational process interventions and cell engineering efforts. The knowledge of the variation in gene expression, can be used for screening cell lines and to identify the host cell most suited to the product being produced.

Chapter 5: Transcriptome dynamics during cell line development in CHO cells

5.1 Context statement

The cell line development process for creating a high producing cell line for recombinant protein production, is very time consuming and labor intensive. A study of the molecular changes in the cells during this process will enable us to rationally engineer or screen the cells which may lead to the shortening of the time scale for cell line development. This chapter presents such a global view of CHO cells transcriptome as they transition from non-secretory host cells to hyper secretors during the process of cell line development.

This project was a collaborative effort between Merck & Co., Bioprocessing Technology Institute and University of Minnesota. YST carried out the experiments at Merck & co. The cell pellets were transported to the lab, where NMJ and HL carried out the RNA extraction. NV extracted the DNA. The RNA samples were shipped to Bioprocessing Technology Institute, where SN hybridized the samples to the microarray chip. NV and HL synthesized the cDNA and conducted the qRT-PCR studies for transcript expression quantification. NV conducted qRT-PCR for genomic DNA copy number quantification. The data from the fed batch cultures were analyzed by NV and HL. NV and HL conducted the all the data analysis for this manuscript. The text and figures for the manuscript was developed by NV and WSH.

5.2 Summary

Dihydrofolate reductase (DHFR) system is used to amplify the product gene to multiple copies in Chinese Hamster Ovary (CHO) cells for generating cell lines which produce the recombinant protein at high levels. The physiological changes accompanying the transformation of the non-protein secreting host cells to a high producing cell line is not well characterized. We performed transcriptome analysis on CHO cells undergoing the selection and amplification process. A host CHO cell line was transfected with a vector containing genes encoding the mouse DHFR (mDHFR) and a recombinant human IgG (hIgG) and subjected to selection and amplification. Clones were isolated following selection and subcloned following amplification. Control cells were transfected with a control plasmid which did not have the hIgG genes. Although methotrexate (MTX) amplification increased the transcript level of the mDHFR significantly, its effect on both hIgG heavy and light chain genes was more modest. The subclones appeared to retain the transcriptome signatures of their parental clones, however, their productivity varied among those derived from the same clone. The transcript levels of hIgG transgenes of all subclones fall in a narrower range than the product titer, alluding to the role of many functional attributes, other than transgene transcript, on productivity. We cross examined gene functional class enrichment during selection and amplification as well as between high and low producers and discerned common features in them. We hypothesize that the role of amplification is not merely increasing transcript levels, but also enriching survivors which have developed the cellular machinery for secreting proteins, leading to an increased frequency of isolating high-producing clones. We put forward the possibility of assembling a hyper-productivity gene set through comparative transcriptome analysis of a wide range of samples.

5.3 Introduction

In the past quarter century, we have witnessed a two order of magnitude increase in product titer (Wurm, 2004). Critical to attaining a high productivity process is the

establishment of a high producing cell line which can secrete the product to a high level in a timely fashion.

Many of the high producing CHO cell lines were obtained using the dihydrofolate reductase (DHFR)-based amplification system developed three decades ago. DHFR catalyzes the conversion of dihydrofolic acid to tetrahydrofolic acid, a precursor for the synthesis of glycine, thymidine phosphate, and purine. CHO cell lines, such as DG44 and DXB11, which are mutated to be deficient in DHFR activity, require supplementation of hypoxanthine and thymidine (HT-media) for growth. Upon co-introduction of the exogenous DHFR gene with the product transgene, cells that express the transgene can be selected in HT-deficient media (Urlaub & Chasin, 1980). By including in the medium a high level of methotrexate (MTX), a DHFR inhibitor, one can further select for cells which have elevated levels of DHFR. The selected cells typically have multiple copies of an approximately 100 kilo base region of the chromosomal region containing the DHFR gene locus (Kim et al, 2001; Kim & Lee, 1999; Milbrandt et al, 1981). This increased copy number of the DHFR gene sequence often increases its transcript and protein levels, enabling them to survive a high concentration of MTX. Since the product transgene is introduced along with the DHFR gene, the product gene is co-amplified with the DHFR gene, leading to an increase in the product gene copy number (Kaufman et al, 1983).

Upon transgene introduction and amplification, parental CHO cells, which do not naturally secrete protein at an appreciable level, can be transformed into secretors. However, the resulting cells often exhibit a wide range of productivity (Jun et al, 2005; Kim et al, 2001; Kim et al, 1998a; Kim et al, 1998b). A screening process thus follows amplification to look for cells which secrete a high level of the product initially; and the selected clones are then subjected to more detailed characterization of production under process conditions.

Despite its widespread use in cell line development over the last quarter century, the transgene amplification and selection of high producing cell lines are still not understood at a mechanistic level. With the arrival of transcriptome analysis and genomic tools, a global survey of the impact of gene amplification on cell line development is called

for. Herein, we report changes in transcriptome during the processes of selection and amplification and deliberate about the functional classes which may contribute to the hyper-productivity trait on a production cell line.

5.4 Materials and Methods

5.4.1 Cell Line Development

5.4.1.1 Transfection and Selection

CHO DUXB-11 cells were maintained at 37^{o}C and 7.5% CO_2 in the MEMα medium with ribonucleosides and deoxyribonucleosides (Gibco#12571) containing 10% v/v Characterized Fetal Bovine Serum (Hyclone#SH30071). The cells were grown to 80% confluence and transfected using the Lipofectamine reagent (Invitrogen, Carlsbad, CA) with 8 μg of an expression vector containing (1) hIgG heavy and light chain cDNAs driven by constitutive Cytomegalovirus (CMV) promoters, (2) hygromycin resistance marker under a Thymidine Kinase (TK) promoter and (3) mouse DHFR gene under a retroviral long terminal repeat (MMTV-LTR) promoter. In addition, a control transfection was performed using the backbone of the expression vector, but without the cDNA of hIgG heavy chain and hIgG light chain.

Following transfection, the cells were sub-cloned in 96-well plates in the MEMα medium without ribonucleosides and deoxyribonucleosides (Gibco#12561) and supplemented with 10% v/v Dialyzed Fetal Bovine Serum (Hyclone#SH30079) and 400 μg/mL hygromycin B. After three weeks, the supernatant of wells containing single-cell colonies was collected, and the hIgG concentration was measured using the enzyme-linked immunosorbent assay (ELISA). Three clones, with highest titers namely P_1, P_2 and P_3, were selected for further analyses. A control cell line (C_0) was also developed, by transfecting similar vectors with all selection marker, but without hIgG cDNA.

5.4.1.2 Methotrexate Treatment and Subcloning

The three clones (P_1, P_2 and P_3) and the control (C_0) were expanded to about 40% confluence in T-75 flasks and further treated with 20 nM methotrexate (MTX) (Sigma Aldrich, St. Louis, MO) for 15 days. The concentration of MTX was optimized by conducting a kill curve on the parental cells. Following MTX treatment, the cells were grown to 80% confluence for sample collection. After MTX treatment, the cell pools obtained from P_1, P_2, P_3 and C_0 were denoted as P_{1M}, P_{2M}, P_{3M} and C_M, respectively. Two biological replicates were used for MTX treatment from each clone. Cell samples for RNA extraction were collected for both biological replicates prior to and following MTX treatment.

Following MTX treatment, the three hIgG producing cell lines P_1, P_2 and P_3, were sub-cloned in 96-well plates using MEMα medium without ribonucleosides and deoxyribonucleosides (Gibco#12561) and supplemented with 10% v/v Dialyzed Fetal Bovine Serum (Hyclone#SH30079) and 400 μg/mL hygromycin B. The antibody concentration in the supernatants of wells with single-cell colonies was measured using ELISA. Two sub-clones each derived from P_{1M} and P_{2M} and five sub-clones derived from P_{3M} were selected for further study.

5.4.1.3 Adaptation of Sub-clones to Suspension and Fedbatch Culture

The sub-clones were adapted to growth in suspension in an in-house modified serum-free Excell ACF CHO medium (Sigma C5467) for two weeks by gradual serum reduction. The cells were then cultivated in fedbatch cultures at 70 mL scale maintained at 37°C and 7.5% CO_2 for 12 days with feeding performed at day 3 and day 6. Cell samples were collected during exponential growth phase (day 4) and stationary phase (day 7) for RNA extraction.

5.4.2 RNA Extraction, cDNA Synthesis, and Hybridization

Total RNA was extracted using the RNeasy kit (Qiagen, Valencia, CA) following the manufacturer's recommended protocols with on-column DNase I digestion to remove any genomic DNA. The quality of the extracted RNA was examined using the Agilent

RNA 6000 Nano chip. cDNA and labeled cRNA were prepared from the RNA extracted for hybridization onto an in-house Affymetrix microarray customized for CHO cells containing 61,223 probe sets representing 26,227 unique gene IDs and 14,657 unique Ensembl mouse genes. cRNA was synthesized from cDNA using the 3'-IVT Express kit (Affymetrix, Santa Clara, CA). Subsequent hybridization onto the custom microarray in the GeneChip Hybridization Oven 640 (Affymetrix, Santa Clara, CA) and processing was done using the GeneChip Fluidics Station 450 (Affymetrix, Santa Clara, CA).

5.4.3 Microarray Data Processing and Analysis

The raw image files were processed using the Expressionist software (GeneData, Basel, Switzerland) for preliminary data filtering. The Affymetrix MAS5 algorithm was used to estimate and correct for background and also to remove and mask defects. The data were then linearly scaled to a mean value of 500 per array. Probe sets with a detection p-value ≤ 0.04 and an intensity ≥ 40 in at least one sample were retained for further analysis.

Hierarchical clustering was performed using the UPGMA (un-weighted average) clustering method in Spotfire DecisionSite (TIBCO, Somerville, MA). Euclidean distance was used to represent the difference in gene expression of the samples under consideration. Significance Analysis of Microarray (SAM) (Tusher et al, 2001) was used to determine the statistical significance of differentially expressed genes. The differential expression of gene classes with related functions (known as gene sets) was analyzed using Gene Set Enrichment Analysis (GSEA) (Subramanian et al, 2005).

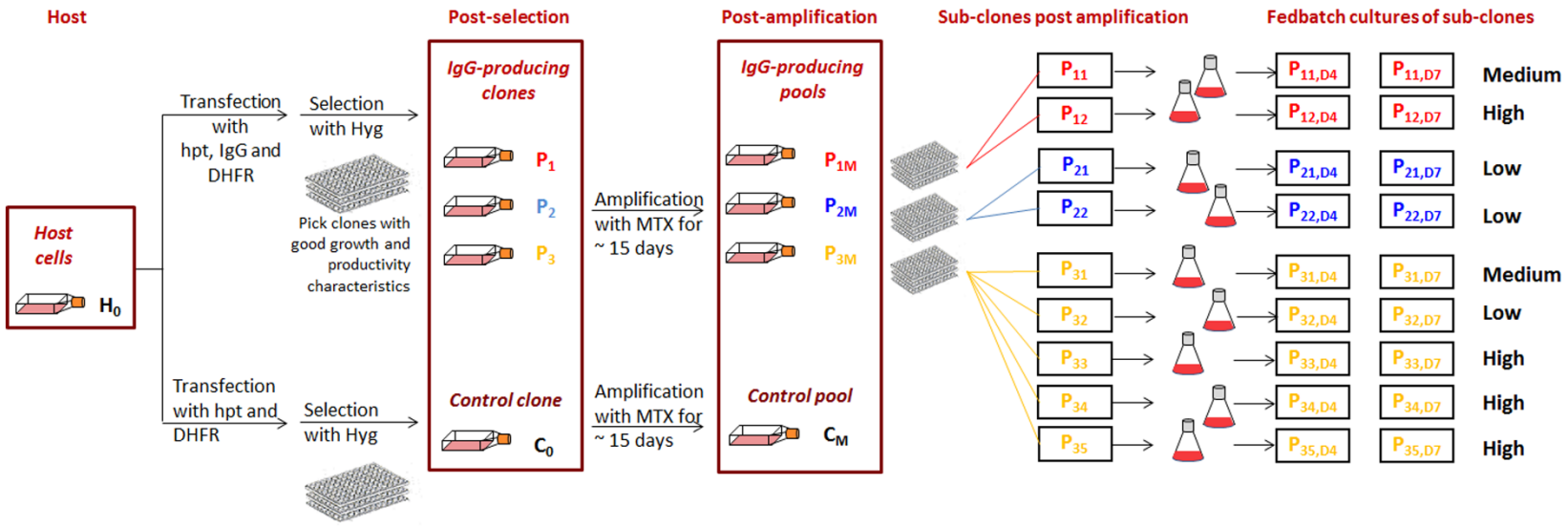


Figure 0-18: Experimental design of selection, amplification, and sub-cloning. Host cells (H0) were transfected with rIgG heavy chain, light chain, mDHFR, and hpt (hygromycin phosphotransferase) genes, and selected in hygromycin and HT-minus media They were single cell cloned to give rise to three clones P_1, P_2 and P_3. A control transfectant (C_0) was developed by transfection of mDHFR and hpt genes into host cells (H_0), and selected in the same media. Each of the three selected clones (P_1, P_2, and P_3) along with the control transfectant (C_0), were subjected to amplification using methotrexate for 15 days. Following MTX treatment, the samples for the three hIgG-producing clones was named P_{1M}, P_{2M} and P_{3M}, respectively. The post-amplification control sample was named C_M. P_{1M}, P_{2M} and P_{3M} were further sub-cloned to give rise to two (P_{11}, P_{12}), two (P_{21}, P_{22}), and five (P_{31-35}) sub-clones, respectively. Each of these sub-clones were cultured in fedbatch mode and assayed for growth and titer. Samples for RNA extraction were collected from day 4 and day 7 of the fedbatch culture.

5.4.4 Quantitative Real Time PCR (qRT-PCR)

qRT-PCR assay was conducted in triplicate using 1 µL of cDNA in a 12.5 µL reaction volume using the Brilliant II SYBR Green master mix (Agilent, Santa Clara, CA) on a MxPro3000P (Stratagene, Santa Clara, CA) machine using standard cycling conditions. hIgG heavy chain, hIgG light chain, mDHFR and beta actin (for normalization) were assayed using qRT-PCR. The primers used for the assay are listed in Table 0-1.

5.4.5 DNA Extraction

5.4.6 Quantification of DNA amplification by qRT-PCR

5.5 Results

5.5.1 Effect of Amplification Process on hIgG Titer and Transcript Level

CHO host cells were transfected with hIgG heavy chain and light chain genes. In addition, a control transfection was carried out using a control plasmid derived from the same construct but without the hIgG genes. Following selection, three clones were obtained in 96 well-plates (P_1, P_2, and P_3) (Figure 0-18). After further cultivation to expand the cells, these three clones were subjected to amplification using MTX. The resulting populations are denoted as P_{1M}, P_{2M}, and P_{3M}, respectively.

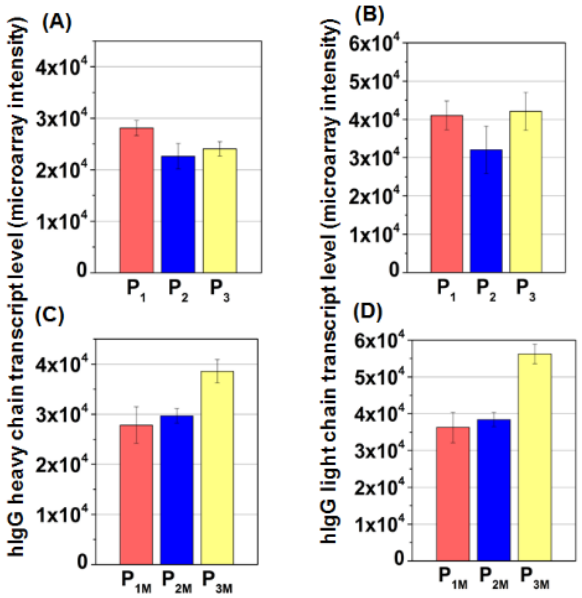


Figure 0-19: Microarray intensity of hIgG upon selection and amplification. (A) Heavy chain upon selection. (B) Light chain upon selection. (C) Heavy chain upon amplification. (D) Light chain upon amplification.

The transcript levels of both the hIgG heavy and light chain genes, after selection and prior to amplification, were rather high, on the order of 10,000 as compared to the mean intensity of 500 for the entire array. Upon amplification with MTX, the transcript levels of the hIgG heavy and light chains in P_1, P_2, and P_3 increased moderately, from an average intensity value of 25,000 to 32,000 for the heavy chain and from an average value of 38,000 to 44,000 for the light chain (Figure 0-19).

To quantify the effect of selection, the transcript level of mDHFR, hIgG heavy chain and light chain were also assayed using qRT-PCR (Figure 0-21). The transcript level of the reference gene beta-actin is shown as a dashed line (i.e., fold-change of 1). mDHFR was expressed at a moderate level upon selection, much lower than that of the reference, beta-actin. In contrast, the expression levels of the hIgG genes were significantly higher post-selection, in agreement with microarray results (Figure 0-19).

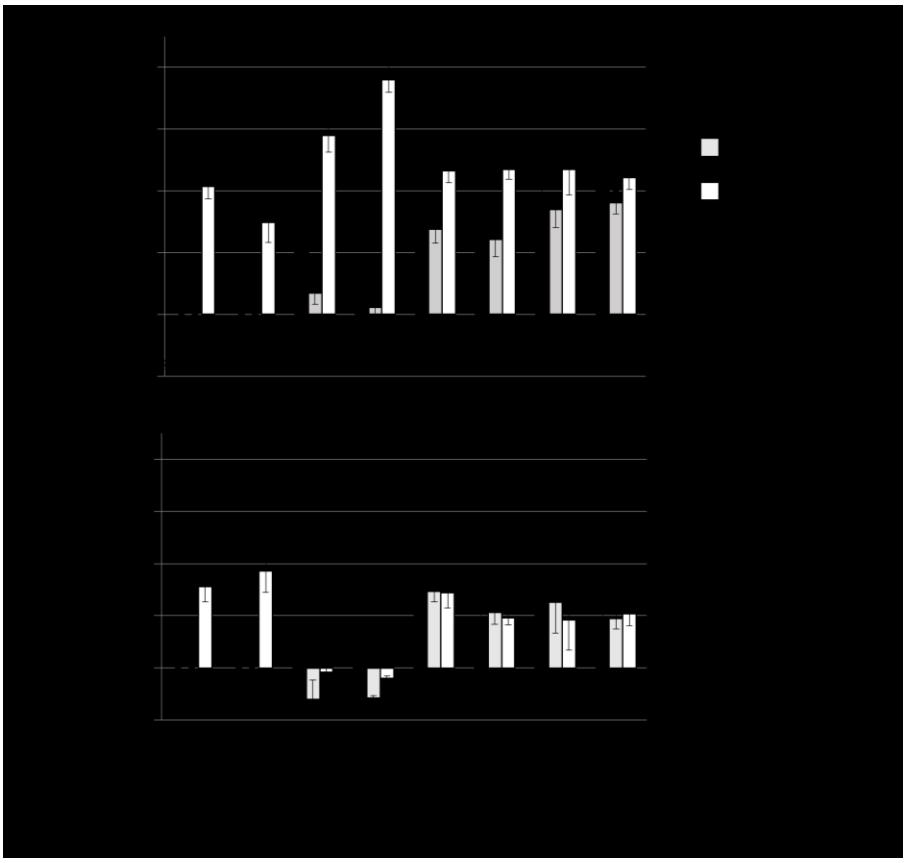


Figure 0-20: Quantification of copy number change of (A) mRNA and (B) DNA post amplification for the control cell line and each of the clones. R1 and R2 are biological replicates.

The amplification process resulted in a 4- to 8-fold increase in mDHFR transcript, which however is still at a level lower than that of beta-actin. In contrast, only a modest change in the transcript level of the hIgG heavy chain and light chain was observed in all three clones P_1, P_2, and P_3 after amplification. However, the hIgG genes were already expressed at higher levels than mDHFR gene, thus even a small fold increase would represent a large number of mRNA molecules (Figure 0-21).

The extent of transgene amplification duplication after MTX amplification for the control clone and each clone and its derived pool is measured compared to its corresponding transcript level increase. After amplification with a moderate level of MTX, a small copy number increase was observed in six pools while two others did not have significant variation. Overall, the extent of increase was higher for mDHFR than for hIgG

genes. The corresponding change in mRNA levels was much higher than the change in copy number of the transgenes, including in the pool that showed no copy number change (Figure 0-20).

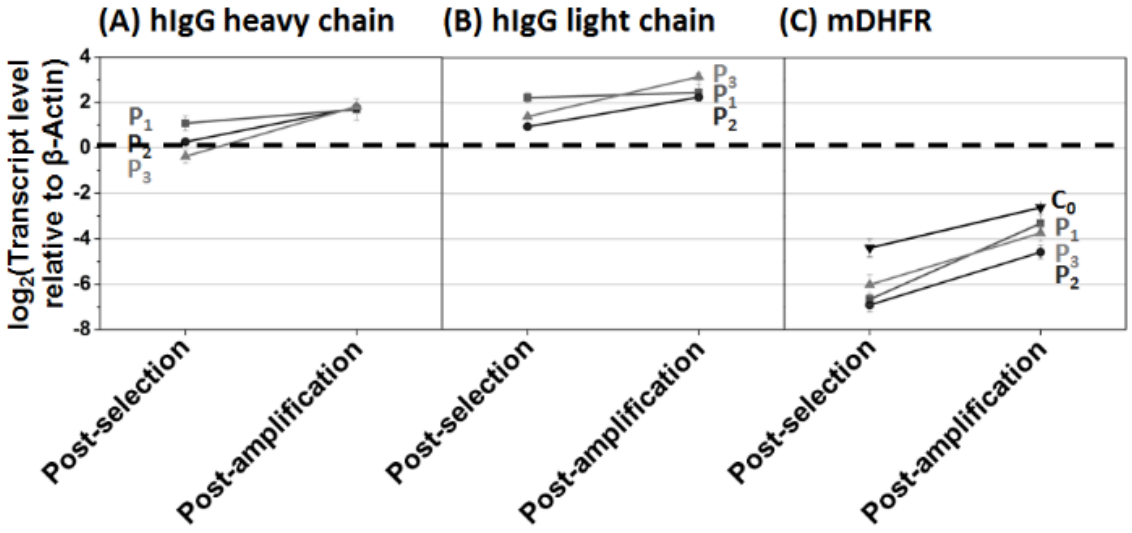


Figure 0-21: Expression level relative to beta-actin upon selection and amplification. (A) hIgG heavy chain. (B) rIgG light chain. (C) mDHFR for clones Control (▼), P1 (■), P2 (▲) and P3 (●).

5.5.2 Characterization of Amplified Sub-clones

A total of nine sub-clones were isolated from P_{1M}, P_{2M}, and P_{3M} (two, two, and five sub-clones, respectively) and denoted as (P_{11}, P_{12}), (P_{21}, P_{22}), and ($P_{31\text{-}35}$) for further analysis. Duplicate fedbatch cultures for each sub-clone were performed and their growth curves are shown in Figure 0-22. Samples from two time points at the exponential and stationary phases (approximately day 4 and 7, shown with dotted vertical lines) of each sub-clone were used for microarray study. mDHFR transcript level was measured by qRT-PCR.

Following MTX treatment and sub-cloning, a wide range of the antibody titer was seen among the sub-clones at the 96-well stage (Figure 0-23A). However, the hIgG concentration measured in the fedbatch cultures at day 4 and day 7 was not clearly correlated to those observed in 96-well plates (Figure 0-23B). Among the three high-titer

sub-clones, P_{12} showed a high productivity in fedbatch cultures while P_{21} and P_{22} did not exhibit high productivity as seen in 96-well culture. Two sub-clones, P_{34} and P_{35}, which had a moderate productivity in the 96-well plate stage, continued to showed a moderate productivity in fedbatch cultures. Notably, two sub-clones, P_{33} and P_{31}, which showed a low titer in 96-well plates, had a moderate productivity in fedbatch cultures. Overall, the final hIgG titer spread over a 4-fold range. A similarly wide span was also observed for specific productivity (Figure 0-23C).

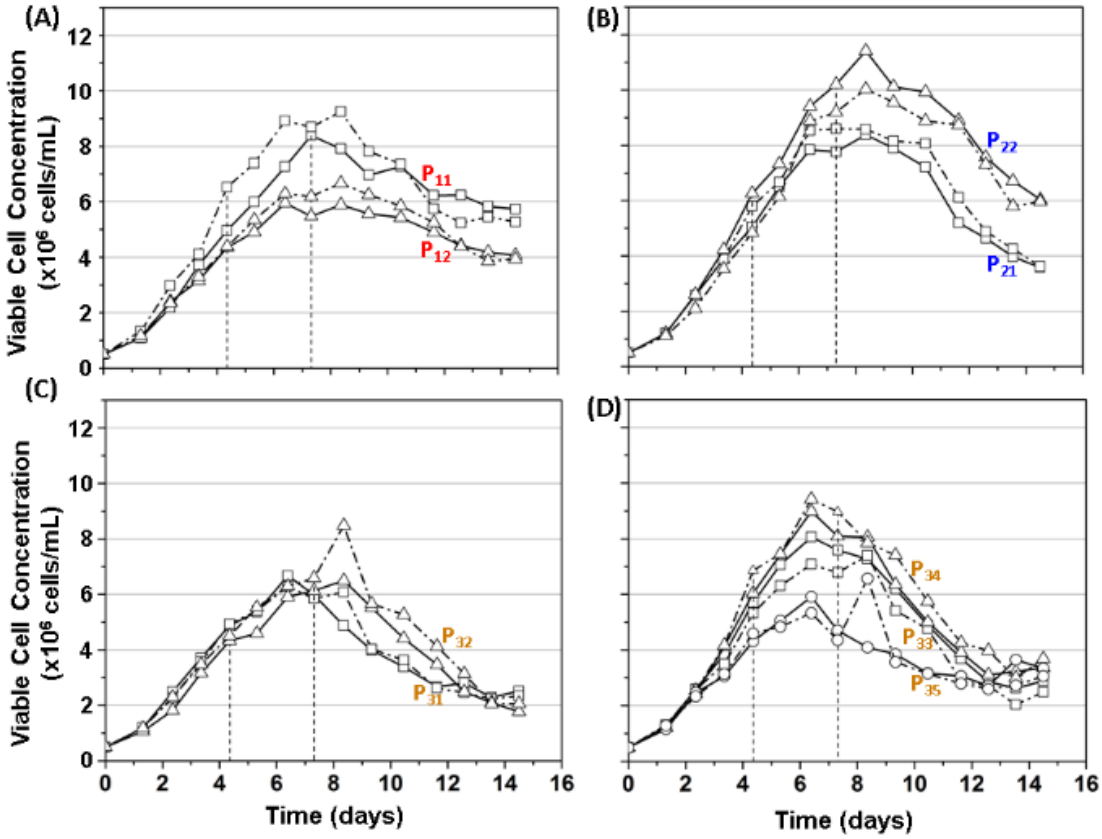


Figure 0-22: Viable cell density in the fed batch cultures of all the nine sub-clones (two replicates each). Replicate 1 is shown in solid line, and replicate 2 is shown in dotted line. Data points for each sub-clone are represented by the same symbol. (A) P_{11} (□) and P_{12} (△). (B) P_{21} (□) and P_{22} (△). (C) P_{31} (□) and P_{32} (△). (D) P_{33} (□), P_{34} (△), and P_{35} (○).

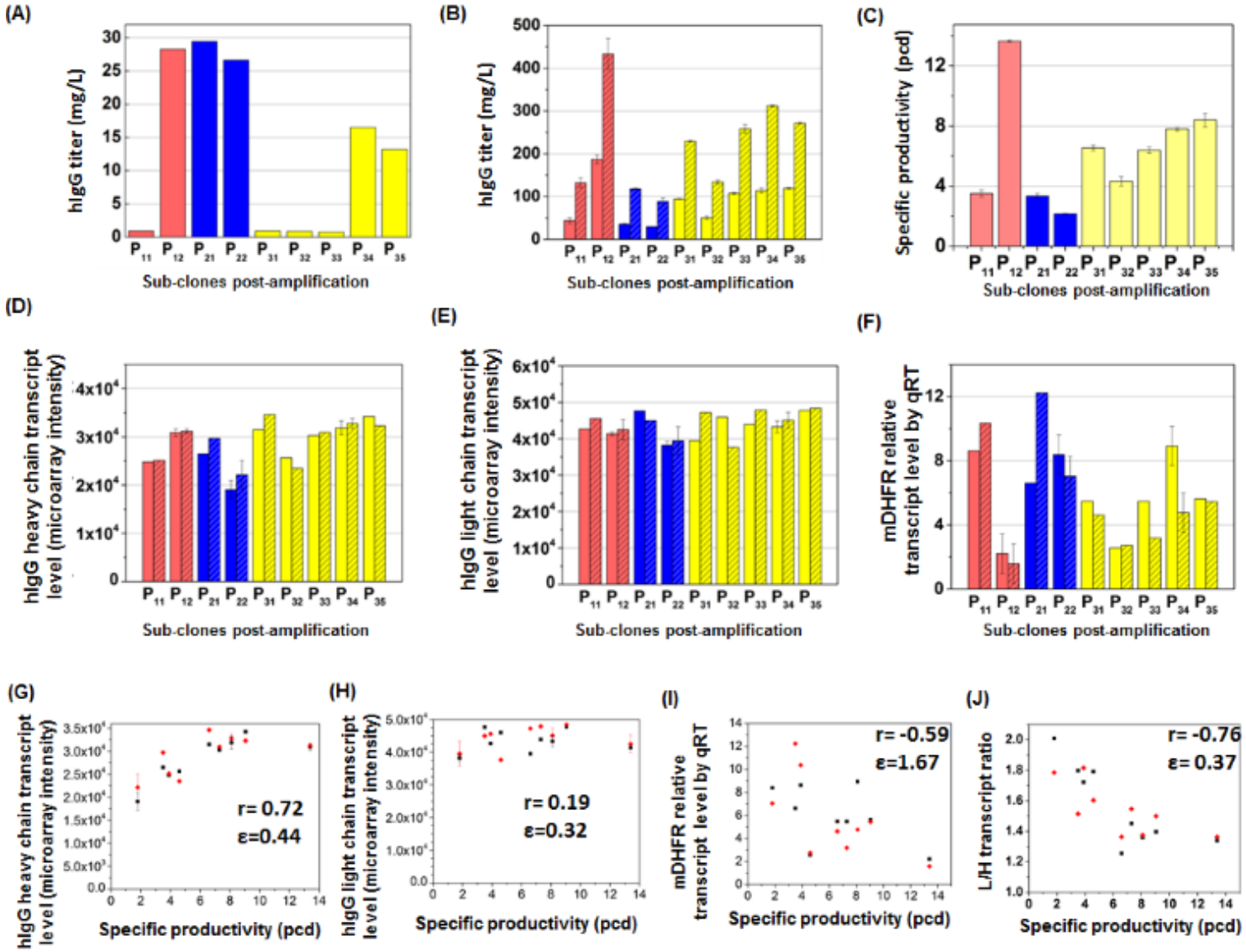


Figure 0-23: hIgG titer in 96-well plate stage, hIgG titer levels in fedbatch culture, specific productivity in fedbatch culture, hIgG heavy chain, hIgG light chain, and mDHFR transcript levels in fedbatch culture are shown in panels (A), (B), (C), (D), (E) and (F), respectively. The solid bars represent day 4 and the dashed bars represent day 7 of fedbatch culture. The Pearson correlation coefficient (r) and normalized root mean square error (ε) for the relationship between the specific productivity and rIgG heavy chain expression, rIgG light chain expression, mDHFR expression, and light to heavy chain expression ratio for all the sub-clones in fedbatch culture are shown in (G), (H), (I) and (J), respectively. Data from day 4 is shown with black symbols and data from day 7 is shown with red symbols.

In contrast to the wide range of productivity, the transcript levels of the light chain and heavy chain hIgG fell into a relatively narrow range for all sub-clones (Figure 0-23D and 5E). The microarray intensities for the heavy and light chain genes spanned from 19,000 to 34,600 and 38,200 to 48,400, respectively. The ratio of light chain to heavy chain transcript ranged from 1.2 to 2, a ratio that has been commonly observed (Schlatter et al. 2005). The mDHFR transcript levels among different sub-clones spanned over nearly an 8-fold range (Figure 0-23F). Compared to hIgG, the variability of the mDHFR transcript levels among different sub-clones was substantially larger.

The transcript level of the hIgG heavy chain showed a positive correlation to specific productivity ($r = 0.72$) (Figure 0-23G). For the hIgG light chain, no apparent correlation between specific productivity and the transcript level was seen ($r = 0.19$) (Figure 0-23H). Surprisingly, the mDHFR transcript level was somewhat negatively correlated to specific productivity ($r = -0.59$) (Figure 0-23I). The ratios of light chain and heavy chain transcripts were also examined, and a negative correlation was discerned ($r = -0.76$) (Figure 0-23J). The root mean square errors (ε) are also shown (Figure 0-23G-J). Except for mDHFR expression ($\varepsilon = 1.7$), the expression trends do not deviate much from the fitted trend ($\varepsilon = 0.3 - 0.4$).

5.5.3 Overall Transcriptome Changes During Selection and Amplification

The three clones and all nine sub-clones isolated were further analyzed for transcriptome profiling. We were constrained in both time and materials necessitating a trade-off between processing more samples with a small set of replicates and fewer samples with more replications for each sample. We opted for a larger number of clones while limiting the number of replicates. For the microarray assays, replicate cultures for all clones were performed. Replicate cultures were also performed for the sub-clones with maximum productivity derived from each clone; while for the rest of the sub-clones, only one of the replicate cultures was assayed.

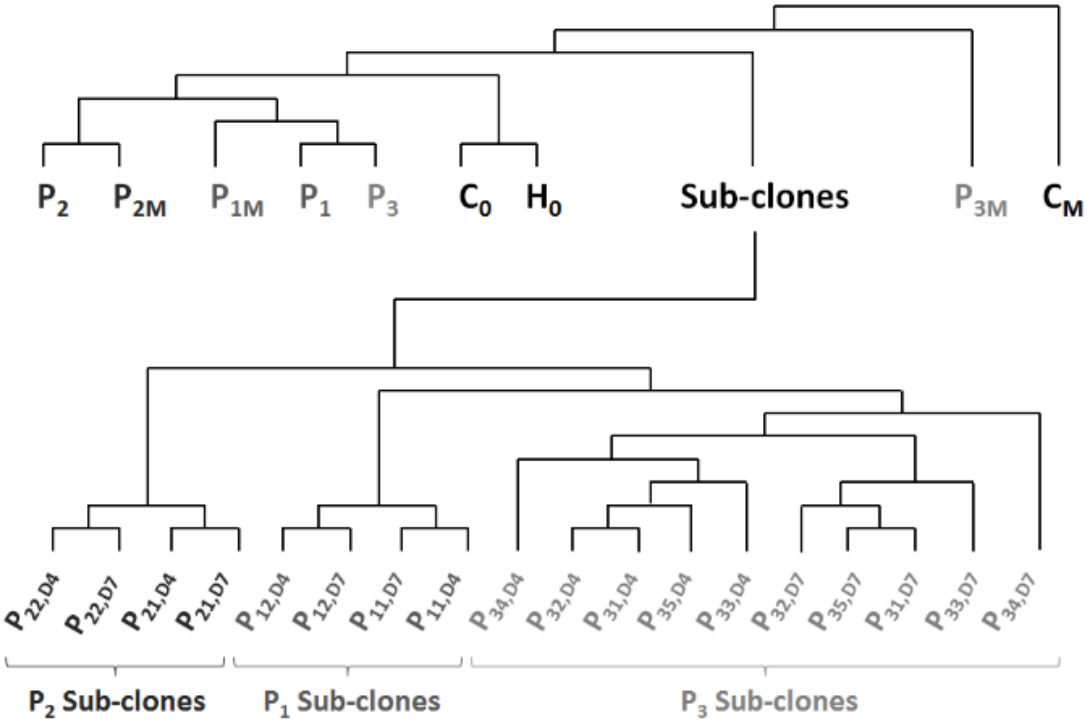


Figure 0-24: Dendrogram from hierarchical clustering of microarray data

5.5.3.1 Investigation of Clonal Variation

Transcriptome data from the host cell and the derived clones and sub-clones (data from fed batch cultures) were clustered using hierarchical clustering (Figure 0-24). The clustering results indicate two major contributions. Among the host and clones, except for P_{3M}, all the other clones clustered together, whereas the host cell and the unamplified control also clustered together. Among all the sub-clones isolated after amplification, those derived from the same clones clustered together. It suggests that significant clonal signature at the transcriptome level is carried forward to the offspring cells regardless of their productivity.

Table 0-2: Generation and enrichment of hyper-productivity gene sets comprising differentially expressed genes from different comparisons.

Process	Selection (S)	Amplification (A)	High vs. Low Producer (H)
Logic	$\left[\left(\frac{\mathbf{P_i}}{\mathbf{H_0}}\right)\ \mathbf{NOT}\ \left(\frac{\mathbf{C_0}}{\mathbf{H_0}}\right)\right]$	$\left[\left(\frac{\mathbf{P_{iM}}}{\mathbf{P_i}}\right)\ \mathbf{NOT}\ \left(\frac{\mathbf{C_M}}{\mathbf{C_0}}\right)\right]$	$\left[\left(\frac{\mathbf{H}}{\mathbf{L}}\right)_{\mathbf{D4}}\ \mathbf{AND}\ \left(\frac{\mathbf{H}}{\mathbf{L}}\right)_{\mathbf{D7}}\right]$
Criteria	$\left(\frac{\mathbf{P_i}}{\mathbf{H_0}}\right): \mathbf{q} < 5$; $\left(\frac{\mathbf{C_0}}{\mathbf{H_0}}\right): \mathbf{q} \geq \mathbf{10}$	$\left(\frac{\mathbf{P_{iM}}}{\mathbf{P_0}}\right): \mathbf{q} < 5$; $\left(\frac{\mathbf{C_M}}{\boldsymbol{C_0}}\right): \mathbf{q} \geq \mathbf{10}$	$\left(\frac{\mathbf{H}}{\mathbf{L}}\right)_{\mathbf{D4}}: \mathbf{q} \leq \mathbf{10}$; $\left(\frac{\mathbf{H}}{\mathbf{L}}\right)_{\mathbf{D7}}: \mathbf{q} \leq \mathbf{10}$
Number of genes after filtering	$\boldsymbol{n(S) = 1996}$	$\boldsymbol{n(A) = 988}$	$\boldsymbol{n(H) = 349}$
Hyper-productivity gene set enrichment			
	Set I	**Set II**	
Criteria	$\boldsymbol{S \cap H}$	$\boldsymbol{A \cap H}$	**-**
Number of genes after enrichment	**58**	**13**	**-**

5.5.3.2 Transcription Variation in Reference to Selection, Amplification, and Productivity

In setting out to construct a gene set pertinent to hyper-productivity, we compared the transcriptome of various cells/clones/sub-clones before and after selection as well as before and after amplification (Table 0-2). The transcriptome of three groups of cells – the host (H_0), the selected clones (P_1, P_2, P_3), and the amplified clones (P_{1M}, P_{2M}, P_{3M}) were

compared using Significance Analysis of Microarray (SAM) (Tusher et al, 2001). The same analysis was also performed on the selected (C_0) and amplified control transfectant (C_M). A q-value of < 5% was used as the criterion to identify differentially expressed genes among producing cells during selection (P_i vs. H_0) and amplification (P_{iM} vs. P_i). These sets of genes were considered responsive to either selection or amplification, but might not be directly related to the production of hIgG antibody. To further limit these gene sets to include only those which are likely to be related to hIgG production, we set out to identify genes which were not significantly differentially expressed in the control transfectant. A q-value criterion of > 10% was used to determine genes not differentially expressed significantly in the comparisons to the control (C_0 vs. H_0 and C_M vs. C_0).

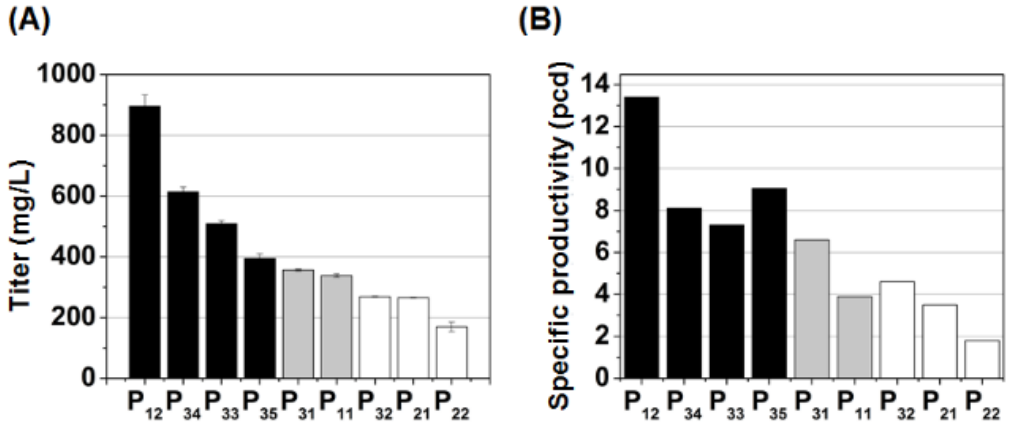


Figure 0-25: The division of high and low producing clones for comparative analysis is shown. (A) shows titer and (B) shows specific productivity of the subclones cultured in fedbatch mode. The dark bars correspond to the sub-clones that are classified as high producers, the white bars correspond to the sub-clones classified as low producers and the grey bars are sub-clones that were mid-producers.

Various sets of genes which were differentially expressed in the producing cell comparison but not significantly differentially expressed in the control were identified: 1996 genes (Set S) are specifically differentially expressed upon selection and 988 genes (Set A) are specifically differentially expressed during amplification. These genes have a high likelihood of being related to the production of hIgG in the course of selection and amplification. The genes which contribute to productivity in high-producing cells have a

high propensity to be in this group. However, not all of these genes are correlated to high productivity because the sub-clones derived from those three producers yielded both high and low product titers.

To further deduce the set of genes which are pertinent to high productivity, we utilized the data which differentiate high- and low-producing cells. The sub-clones fall into three groups (high producers: P_{12}, P_{34}, P_{33}, P_{35}; middle producers: P_{31}, P_{11}; and low producers: P_{32}, P_{21}, P_{22}) in terms of their titer (Figure 0-25A) and specific productivity (Figure 0-25B) observed in fedbatch cultures. The distinction between high and low producers was made such that the high producers show both high titer and high specific productivity. Four and three sub-clones were assigned into high producer and low producer groups respectively. Two sub-clones that do not have consistent pattern in their titer and specific productivity were not included in the high productivity gene set analysis. SAM was performed on fedbatch culture data of the sub-clones of high-productivity and low-productivity groups. Genes differentially expressed between high- and low-producers in day 4 and day 7 of fedbatch cultures were identified. 349 genes were consistently differentially expressed in both day 4 and day 7 with a q-value of less than 10%; these genes were denoted as Set H.

Genes that are common between Set S and Set H possibly represent the high productivity genes which attain their expression levels through selection (Set I). Those in the intersection of Set A and Set H are possibly related to hyper-productivity and are accorded in the amplification (Set II). A total of 58 and 13 genes are in those two gene sets, respectively.

5.5.3.3 Functional Analysis of Differential Expression

To understand the physiological functions possibly conferred by the changes in gene expression in the course of selection and amplification, Gene Set Enrichment Analysis (GSEA) (Subramanian et al, 2005) was performed on the set of data used for the identification of differentially expressed genes. The gene sets used for the analysis were obtained from the Molecular Signatures Database, Broad Institute (v2.5 MSigDB). The

Table 0-3: Functional enrichment during selection, amplification, and high vs. low producers in exponential and stationary phase by Gene Set Enrichment Analysis (GSEA).

Gene sets changed during	Selection		Amplification		High vs. Low Producers			
					Exponential Phase		Stationary Phase	
	Direction	p-value	Direction	p-value	Direction	p-value	Direction	p-value
Signaling Pathways								
EDG1 signaling pathway	Up *	0.015 *	Up *	0.029 *	Down	0.0385		
PDGF signaling pathway	Up *	0.047 *	Up *	0.042 *				
mTOR signaling pathway	Up *	0.055 *						
Toll-like receptor signaling pathway	Up	0.039						
Cytokine-cytokine receptor interaction	Up	0						
RAS pathway			Up	0.069				
Extracellular matrix receptor interaction			Up *	0.048 *	Down	0.0045	Down	0
PIP3 signaling							Up	0.064
IGF1R pathway					Down	0.0042	Down	0.089
TNFR1 pathway					Down	0.0346	Down	0.02
JAK STAT signaling pathway					Down	0.0188		
Gene Expression and Protein Synthesis								
Aminoacyl tRNA biosynthesis	Down	0	Down	0.020	Up	0.0213		
mRNA processing	Up *	0.068 *						
Ribosome	Up	0.080			Up	0.0593		
Proteasome			Down	0.003				
Cell Cycle								
Cell cycle	Up	0.05	Down	0	Down	0.0056	Down	0
DNA replication	Up	0.008	Down	0	Up	0.0076		
CDK5 pathway					Up	0.0503		
Energy Metabolism								
Glutathione metabolism	Down *	0 *						
Amino acid metabolism	Down	0-0.081	Down	0.003-0.080	Up	0.0098-0.0846		
Glycolysis and TCA cycle			Down *	0 *				
Mitochondria							Down	0.04
Others								
Cytoskeleton function					Down	0.0165	Down	0.003
ABC transporters					Down	0.035	Down	0.053

gene sets that were significantly enriched (nominal p-value < 0.1) after selection or amplification as compared to the control were identified. The enriched gene sets fall into two categories: one is enriched in both producing clones and control transfectant, the other is enriched in producing cells but not in the control transfectant. The identified enriched gene sets are listed in Table 0-3. These gene sets are likely to be related to the production and secretion of the product protein since they were enriched only in producing clones but not in the control transfection. We then proceeded to perform GSEA on transcriptome data of high and low producing cells to identify gene sets enriched in the high-producing sub-clones as compared to the low producing sub-clones. These gene sets represent biological functions that are likely to contribute to enhancing productivity.

Interestingly, there is a significant overlap between gene sets which are enriched in the process of cell line generation (i.e., upon selection and amplification) and in productivity enhancement (i.e., high vs. low producing cells), suggesting that high-producing cells may have more enhanced enrichment than the low producing cells in functions which enable a cell to produce and secrete products. Prominent among the gene sets that were enriched between the transfected clones and the parental host during selection were the EDG1, PDGF, mTOR, Toll-like receptor, and Cytokine signaling pathways. Of these, the EDG1, PDGF pathways were also up-regulated following the amplification process, as were Ras, and extracellular matrix (ECM) receptor interaction. Interestingly, the EDG1 pathway and ECM receptor interaction pathways were down-regulated in high producing clones compared to low producing ones. Most of these pathways contain genes from the MAPK signaling pathway (e.g., Mapk1, Mapk3), the PI3K signaling pathway (e.g., Pik3ca, Pik3r1), and the AKT signaling pathway (e.g., Akt1, Src, and Smpd1). Furthermore, several other signaling pathways, including IGF1R, TNFR1, and JAK-STAT, were down-regulated in high vs. low producers whereas the PIP3 signaling pathway was up-regulated under such comparison.

Aminoacyl tRNA biosynthesis and amino acid metabolism were enriched following selection and amplification as well as in high vs. low producers although the direction of change was not the same. Genes involved in mRNA processing and ribosomes

were up-regulated following selection, but unchanged by amplification. The ribosomal elements were further up-regulated in high vs. low producers. Proteasome activities were down-regulated upon amplification and unchanged under other conditions, whereas amino acid metabolism was up-regulated in high vs. low producers.

Genes involved in cell cycle, DNA replication, and CDK5 pathways also showed various enrichment trends upon amplification, selection, and comparison of high vs. low producers. Glutathione metabolism was down-regulated during selection in the producing clones but unchanged in the control transfectant. Glycolysis and TCA cycle were unchanged during selection but down-regulated following amplification. Other gene sets which were enriched in high vs. low producers included cytoskeleton function, mitochondria, and ABC transporters.

5.6 Discussion

Through DHFR-based transgene amplification a non-secretory host cell can be transformed into a hyper-producer. We employed transcriptome analysis to gain insights into the mechanism of this process. A relatively low concentration of MTX was used for a moderate degree of amplification. This protocol was adopted to minimize the emergence of a large number of unstable clones seen when a high MTX concentrations was used (Fann et al, 2000; Kaufman et al, 1985; Pendse et al, 1992). After amplification, the transcript level of mDHFR increased about 4-fold. However, the change in the level of hIgG heavy chain and light chain was only moderate (Figure 0-21A and 3B). This moderate change was still seen in the sub-clones after single cell cloning (Figure 0-23D and 5E). Given the long held notion that amplified gene copy number would give rise to higher transcript level, thereby enhancing productivity, this absence of a large increase in the transgene transcript level was surprising.

The small change in copy number of transgene is consistent with the low concentration of MTX used in this study. The amplification thus only caused a rather moderate change at the transgene copy number level (Figure 0-20B). However, the change in transcript levels, was much higher than the change in the copy number (Figure 0-20A).

Thus, increasing the transgene copy number is potentially not the only role of MTX amplification in selecting hyper productive cells.

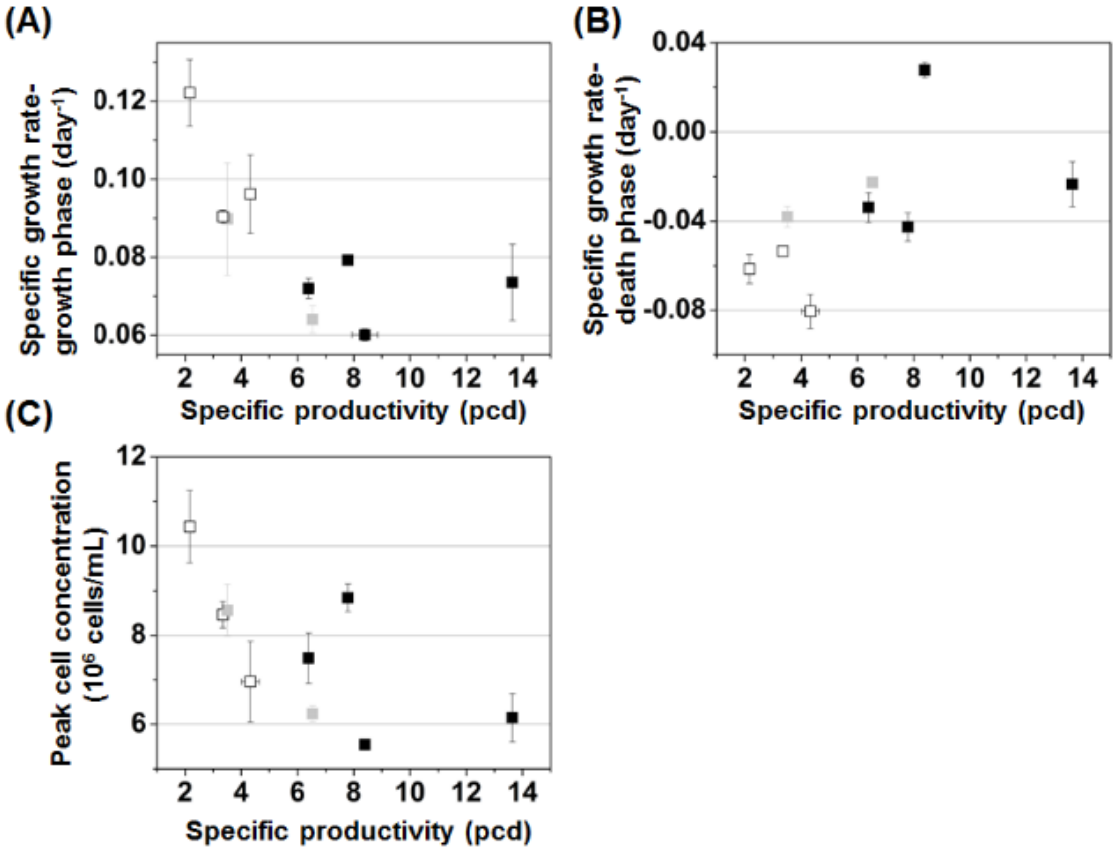


Figure 0-26: Scatter plot showing correlation between specific productivity and (A) specific growth rate in growth phase, (B) specific growth rate in death phase and (C) peak cell concentration in fedbatch culture for high-, low- and mid-producers. The (■) symbols correspond to the sub-clones that are classified as high producers, the (□) symbols correspond to the sub-clones classified as low producers and (▩) symbols are sub-clones that were mid-producers.

Although there was only a small variation in expression level of IgG genes among the sub-clones, a wide range of specific productivities was observed among them. A weak positive relationship between productivity and the transcript level can be seen in heavy chain, but not in light chain (evidence also shown in (McLeod et al, 2011)). The lack of a strong direct correlation between the transgene transcript level and the product titer has also been observed before (Fann et al, 2000; Flickinger et al, 1992; Leno et al, 1992). All

clones and sub-clones had a light chain transcript to heavy chain transcript ratio larger than 1, as has been reported for both transcript and protein levels (Bibila & Flickinger, 1991; Li et al, 2007; O'Callaghan et al, 2010; Schlatter et al, 2005). However, the high producing sub-clones only had a smaller degree of excess level of light chain transcript, while low producers tend to have a light chain to heavy chain transcript levels above 1.5.

High producing subclones show lower specific growth rate and lower specific death rate (Figure 0-26A & B). However, the data set is insufficient for asserting a firm correlation between growth rate and productivity. Two fast growing sub-clones which also gave high maximum cell concentrations were both lower producers (Figure 0-22). It is worth noting that similar observations of a weak correlation between growth behavior and productivity characteristics have been reported previously (Chusainow et al, 2009; Fann et al, 2000). Functional analysis also revealed that the cell cycle progression functional class is significantly down-regulated in high producers compared to low producers (Table 0-3).

The maximum cell concentrations seen in the high producing clones spanned over a range, but did not bear any relation to the productivity of the subclones (Figure 0-26C). In this study and in industrial practice, a high producer is judged by a high product concentration at the end of production. The product titer is affected by both growth and specific productivity. Thus, it is not surprising that among clones of similar product titer, a range of growth behavior is seen.

Hierarchical clustering of transcriptome data showed that, except P_{3M}, all hIgG producing clones (P_1, P_2, P_3) and amplified subclones cluster together, suggesting that the production and secretion of IgG (as a result of selection) has a profound effect at the transcriptome level that is discernable in all secreting clones. Sub-clones derived from the same parental clone retained similar transcriptional signature and clustered together. Sub-clones of similar productivity did not show a high similarity in clustering. The results thus suggested that the hyper-productivity related gene expression profile may not be sufficiently distinctive in genome wide transcriptome analysis.

It was postulated previously that the hyper-productivity is likely to entail relatively small expression level changes in a large number of genes in many functionalities, rather

than profound and readily discernible alterations in a small set of genes (Seth et al, 2007b). Different high producing clones may acquire the same characteristics through different gene expression alterations.

We next focused on those genes whose expression was significantly altered in the producing cells during selection and amplification but not in the control. It is not surprising that the number of genes identified (sets I and II) is relatively small. We employed Gene Set Enrichment Analysis (GSEA) (Subramanian et al, 2005) to compare the functional classes that have been enriched during the process of selection and amplification, as well as between low and high producing cells. In performing GSEA genes with related biological significance, such as the same pathway or imparting the same consequence (e.g. correlated to the occurrence of a particular disease) are grouped into "gene sets". By comparing the transcript levels of genes in a gene set of two groups of samples, GSEA makes a statistical call whether the changes (called enrichment) in this gene set is above that caused by random events. The enrichment call is thus affected by the gene composition in a gene set.

A number of gene sets enriched in selection and amplification are related to growth control, including mTOR, PDGF, and RAS signaling pathways (Table 0-3). Given the toxicity of selective markers used in selection and amplification, it is not surprising that growth control-related functional classes were enriched. This was consistent with the enrichment of cell cycle related functional classes, also observed in another transcriptome study (Doolan et al, 2013). The enrichment of the general functional classes related to cell cycle, cytokine signaling, energy metabolism and redox balance (e.g., glutathione metabolism) finds parallel in the gene expression changes seen in the development of non-secreting B cells to immunoglobulin secreting plasma cells (Bertolotti et al, 2010; Underhill et al, 2003; Vene´ et al, 2010). The detailed list of genes is presented in the Appendix Table 0-2.

This study represents a first attempt to systematically study the mechanism behind the transformation of a non-secretory host cell into a high producer of recombinant protein. A surprised finding is the modest change of transgene transcript level imparted by

amplification, making one wonder the role of amplification. A possible explanation is that the high transcript level of the transgene after selection may have surged to an even higher level during the 15 day exposure of high level of MTX for amplification, which would have caused a surge in the translation of IgG. As a result, cells with an insufficient secretory capacity would have excessive accumulation of intracellular product protein and encountered death. Only those which develop the necessary machinery to allow for high level of IgG synthesis and secretion would have survived. Hence, the amplification process increases the frequency of the isolation of hyper-producing clones. The unfolded protein response (UPR) pathway may play an important role in this process (reviewed in (Hussain et al, 2014; Prashad & Mehra, 2014)).

The small number of cells in culture and the extensive cell death during amplification does not permit a comprehensive transcriptome analysis. The transcription array assay was performed on the amplified cell pools after a period of cell recovery. It did not capture the transient dynamics at the transcriptome level. Nevertheless, the results of GSEA analysis on the selection and amplification process revealed the enrichment of signaling pathway, aminoacyl tRNA synthesis, mRNA processing and ribosome machinery. These changes may be the remnants of the amplification process.

The process of selection and amplification would have set of gene expression profile changes related to increased productivity. The transcriptome comparison between high and low producers would also have highlighted those genes entailed in increasing the productivity. The results of our attempt of integrating those two transcriptome analysis indicate that the road to hyper-productivity does involve colossal changes in gene expression across a wide range of functional classes. This is largely consistent with other reported transcriptome analysis on recombinant CHO (Grillari et al, 2001; Nissom et al, 2006) and mouse myeloma cells (Charaniya et al, 2009; Seth et al, 2007b) of different productivities as well as on recombinant CHO cells under culture conditions which enhance productivity (Kantardjieff et al, 2010a; Yee et al, 2008; Yee et al, 2009).

GSEA studies on cell lines of bioprocess interest, have provided some insight, however, the results have not been very revealing. The gene sets employed were based on

biological function classes, or were collections of genes which were reported to be related to a disease (such as a cancer) or a biological event (such as a developmental process). Ideally, genes which have a high probability of being differentially expressed between high producing and low producing cells or host cells should be collected as a hyper-productivity gene set. An enrichment of such a gene set will be indicative of a clone's propensity to become a high producer during cell line development.

One may hypothesize that the generation of a hyper-producing cell line and the culture conditions which increases the productivity share similar pattern of gene expression changes. Furthermore, such hyper-productivity features may be shared across cell line derived from different tissues and different species, such as CHO and mouse myeloma cells. With that premise one may proceed to assemble a hyper-productivity gene set by comparative analysis of transcriptome data from different sources.

The assembly of a hyper-productivity gene set will certainly require a large number of cell and culture samples across a wide range of conditions. The data presented in this study are very small in comparison to what is needed; considering that alternative genes may give rise to the same trait, the sets of genes generated in this study may be only tentative. Nevertheless, we compared the combined list of the genes which were reported to be differentially expressed in previous transcriptome studies (Charaniya et al, 2009; Kantardjieff et al, 2010a; Seth et al, 2007b; Yee et al, 2008; Yee et al, 2009) with the genes identified from our analyses. From among a total of 451 genes reported in these previous studies, 15 were identified as differentially expressed in the high vs. low producing sub-clones (Set H) in this study. These 15 genes have function related to cell cycle (Ncapd2, Tfdp1), regulation of transcription (E2f6), chromatin organization (H2afy), lipid metabolism (Hmgcr, Acs14, Hsd17b12), protein processing (Dnpep, Fn1, Timp2), protein secretion (Arfrp1, Ckap4), protein synthesis (Rps20) and signaling pathways (Sar1b, Pask) (Listed as Set III in Table 0-2). A gene set of 15 genes is too sparse, while that of 451 is too large (Subramanian et al, 2005). With the increasing availability and affordability of RNA seq analysis, the data set available will increase and the establishment of a hyper-productivity gene set may be feasible.

Chapter 6: Engineering dynamic nutrient uptake for improved CHO cell culture performance

6.1 Context statement

This chapter demonstrates the utility of a dynamic promoter which enables the expression of a transgene in a dynamic fashion. The transgene exhibits low expression in the lag and exponential growth stages, but a high expression in the stationary phase. Such a dynamic promoter was identified by mining a large amount of time-series microarray data. The microarray data was collected and processed by AK. HL reorganized and analyzed the data to identify dynamic genes. HL isolated the dynamic promoters from Chinese hamster genomic DNA and conducted verification in GFP system. NV constructed plasmids for mGLUT5 overexpression and constructed stable cell lines for experiments. All the characterizations in fed batch culture were conducted by NV with the assistance of XZ. Data analysis was conducted by NV. The portions to which I have contributed significantly have been retained. The remaining sections were condensed and replaced with a summary to provide adequate context for the following sections.

6.2 Summary

Cell engineering efforts in mammalian cells are conventionally performed using constitutive expression systems. However, cells respond to various environmental cues and cellular events dynamically according to the cellular needs. The use of inducible systems allows for time dependent expression, but it also requires external manipulation. Ideally, a transgene's expression should be synchronous to the host cell's own rhythm, and its expression level should be adjusted to the cellular process and manipulated according to

the transgene's function. To that end, we surveyed the transcriptome profiles over different stages of fed-batch cultures operated under different conditions and identified genes with different expression dynamics and intensity ranges. The promoters of these genes are likely to drive the expression of the transgenes following patterns similar to those of the endogenous genes. To illustrate their potential utility for conferring dynamic responses, a representative promoter of the Thioredoxin-interacting protein (Txnip) gene was used to drive expression of a blasticidin resistance (BSD) gene and an enhanced green fluorescent protein (EGFP) gene in concert with cell growth. We further employed this Chinese hamster promoter to engineer dynamic expression of the mouse GLUT5 fructose transporter in Chinese hamster ovary (CHO) cells, enabling them to utilize sugar according to cellular needs rather than in excess as typically seen in culture. Under these conditions, less lactate was produced, resulting in a better growth rate, which in turn prolonged culture duration, and increased product titer. 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.

6.3 Introduction

A producing cell line is developed by the introduction of a transgene encoding for the protein product into the host cells, usually followed by gene copy number amplification to increase its transcript level (Kantardjieff et al, 2010b; Seth et al, 2006a). Subsequently, extensive screening and testing are performed to isolate producing cell lines that give rise to high levels of the product in the production culture conditions (Freshney, 2005; Hu, 2012; Wirth & Hauser, 2008). However, high transcript level of the product gene is not sufficient for high productivity and consistency of product quality. The combination of many superior characteristics that collectively confer the complex trait of hyper-productivity (Chapter 5 and (Meleady et al, 2011; Seth et al, 2007a)). In cell line development, a large effort is typically devoted to screen for cell clones which confer major attributes which collectively give the desired characteristics of a hyperproducer.

A possible faster route for cell line development is via cell engineering to endow cells with the desired characteristics, especially on host cell lines that can be used for product transgene introduction. The engineering of metabolic, secretory, and growth control pathways in producing cells to enhance their growth and product secretion trait have all been attempted (reviewed in (Datta et al, 2013; Kim et al, 2012) (Kaufmann & Fussenegger, 2003; Seth et al, 2006a; Vallee et al, 2014)). For instance, down-regulation of lactate dehydrogenase A (LDH-A) and pyruvate dehydrogenase kinases (PDHK) to reduce lactate production (Zhou et al, 2011); over-expression of the transcription factor X-box binding protein 1 (Xbp1) to enhance secretory capacity (Tigges & Fussenegger, 2006); and over-expression of the anti-apoptotic genes E1B19K and Aven to delay the onset of apoptosis (Figueroa et al, 2007) have all been reported.

The expression of these genes is typically driven by a constitutive promoter, such as those derived from cytomegalovirus (CMV), Simian virus 40 (SV40), or elongation factor 1 α (EF1α). Such strong and constitutive expression may not be suitable for all genes (Figure 0-27). In some cases, it is desirable to express the transgene only in a particular time or condition of the culture, especially when the gene confers a negative effect, such as growth inhibition, on cellular behavior. The use of inducible promoters allows for possible dynamic expression to mitigate the negative effects. However, depending on the inducible system employed, external addition of an inducer such as tetracycline, isopropyl-β-D-thio-galactoside (IPTG), or a hormone is necessary. The use of such inducers, especially antibiotics, is undesirable at a manufacturing scale. 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 (Figure 0-27).

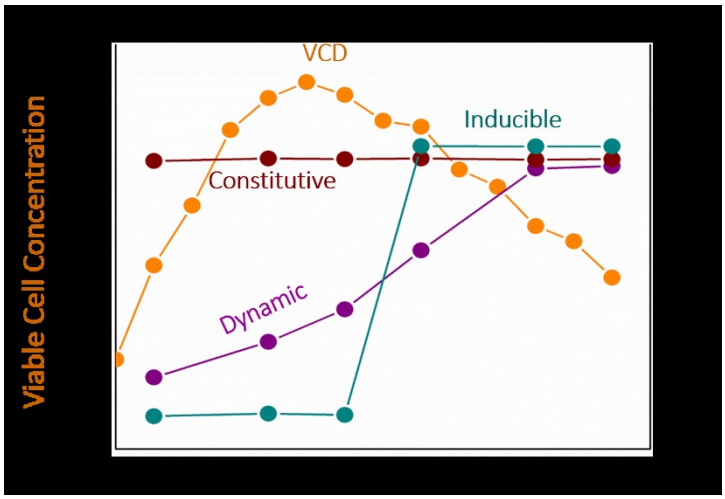


Figure 0-27: Schematic of transgene expression patterns in a typical fed batch culture of engineered mammalian cells. Viable cell density curve shows exponential phase stationary phase and death phase of the fed batch culture. Most cell engineering is constitutive where the transgene is expressed at continuously high levels throughout the course of the culture. In contrast, inducible cell engineering strategies, make use of an inducer molecule to induce gene expression at a desired stage of the culture, and the high expression level is maintained for the period that the cells are exposed to the inducer. Dynamic cell engineering involves using promoters whose natural transcriptional response is dynamic.

The endogenous promoter can be used for similar cell engineering strategies, especially in engineering the metabolism of mammalian cells because such engineering often requires a dynamic control on gene expression. For example, it is well known that cellular metabolism is very different in the growth and stationary stages of the culture (Wahrheit et al, 2014). Lactate accumulation is major obstacle in achieving high productivity, and hence, higher titers (Glacken et al, 1986; Hu et al, 1987). However, high lactate is essential for good growth (reviewed in (Mulukutla et al, 2010)) akin to the Warburg effect in cancer cells. Therefore, an ideal process would involve high lactate accumulation in the initial culture stages to boost growth, and lower lactate concentrations at the stationary stage to enhance productivity. Data mining studies on industrial bioreactors have indicated lactate consumption phenomenon to be the most correlated with better reactor yields (Le et al, 2012). This is usually achieved by introducing metabolic shifts, i.e., inducing the consumption of lactate at the stationary stage to lower lactate

concentrations. Metabolic shifts can be introduced by lowering glucose concentration in the stationary phase of the culture to extremely low levels, ~ 0.04 g/L (Cruz et al, 1999; Mulukutla, 2012; Zhou et al, 1997) in order to maintain low glycolytic flux. Since the glucose transporter glut1 has a high affinity for glucose, the enzyme attains saturation kinetics at very low glucose concentrations (Figure 0-28). Maintaining such low glucose levels in large bioreactors is difficult because they require precise control of glucose feeding with a continuous estimate of glucose concentration. In addition, insufficient in mixing can create localized pockets of starved cells where the glucose concentration can drop quite rapidly at high cell densities.

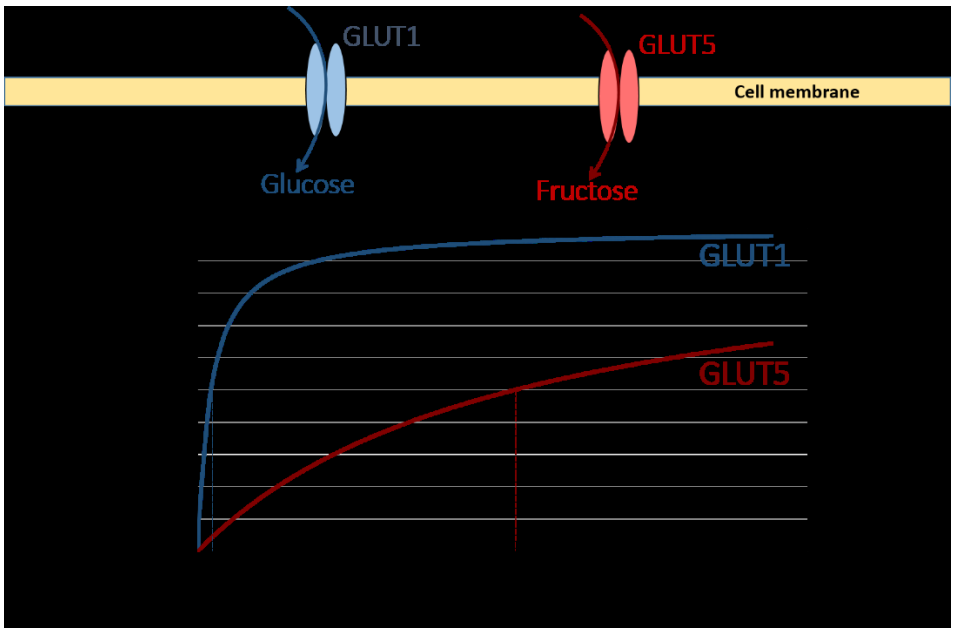


Figure 0-28: Kinetic behavior of sugar transporters *glut1* and *glut5*. *Glut5* preferentially transports fructose and has a much higher K_m value for fructose compared to the K_m value that *glut1* has for glucose. K_m values are indicated by the dashed vertical lines.

Cell engineering can provide a solution to this problem. One can knockdown the glucose transporter conditionally in the late stage of the culture to enforce lower transport flux. However, the cells overexpress other glucose transporters glut4 and glut8 to overcome this restriction (Bengea, 2008). Another possibility is to switch the sugar source to a substrate that is not taken up by the glucose transporters. For example, fructose is taken up

by the *glut5* transporter, which has a much lower affinity for fructose than that glut1 has for glucose (Figure 0-28). Since, glut5 is not expressed in CHO cells, they can easily be engineered to consume fructose by knocking in the *glut5* gene. Moreover, a dynamic promoter can be used to express the *glut5* gene to ensure low expression in the growth stage, and higher expression in the stationary phase when lower glycolytic flux needs to be engineered.

The identification of genes and their promoters with the desired expression dynamics necessitates the use of time profiles of transcript levels to discern unique expression patterns. Major expression trends of hundreds of regulatory motifs or genes were distinguished from the background noise either by model fitting or by principal component analysis (PCA). These trends could be subsequently classified into several groups by clustering based on their similarity.

The next challenge to be addressed is the isolation of the corresponding promoter region because the precise locations and boundaries of the DNA binding region of regulatory elements are not easily identifiable. Most mammalian core promoters are associated with either a TATA box or a CpG island within a few hundred basepairs from the putative transcription start sites (Carninci et al, 2006). The proximal promoter regions, however, can extend to several kilo basepairs and comprise binding sites of different transcription factors, many of which are not known. The core promoter often contributes to the basal expression level of the downstream gene, whereas different regulatory elements in the proximal promoter region typically confer transcriptional responses of that gene to various stimuli. Due to this fuzzy boundary of promoters, a relatively long fragment of at least one thousand basepair is often isolated (Minn, 2005; Thaisuchat et al, 2011; Wang et al, 2005). Serial deletion and other modifications are performed subsequently to identify *cis*-regulatory elements within the isolated promoter fragment.

A survey of historical time-course microarray data from multiple fed-batch cultures were used to identify genes with time dynamic expression trends in CHO cells. The promoter of a candidate gene (Txnip) was used to drive the expression of a fusion gene

(BSD_EGFP) following the expected dynamic manner. Furthermore, dynamic expression of the GLUT5 fructose transporter under the control of this Txnip promoter resulted in more moderated sugar consumption, less lactate production, better growth characteristics, and higher product titer. Thus this study represents the first proof-of-concept in dynamic cell engineering.

6.4 Materials and Methods

6.4.1 Construction of Expression Vectors

The pTet_BSD_EGFP plasmid with the tetracycline inducible promoter, pTet, in this plasmid replaced by the Chinese hamster Txnip promoter via the restriction sites of AatII and EcoRI to generate the pTxnip_BSD_EGFP plasmid (Figure 0-29B). A similar vector, pCMV_BSD_EGFP, wherein the human CMV promoter was used to drive the expression of the fusion gene, was created similarly using sequential digest of NaeI and EcoRI. The mouse GLUT5 cDNA was amplified from the pcDNA3.1-GLUT5 plasmid (Wlaschin & Hu, 2007a) using the two primers GLUT5_L (GTACCGAGCTCGGATCCACTAGTCC) and GLUT5_R (AACGGGCCCTCTAGAATCGATCGGCCGCCC).

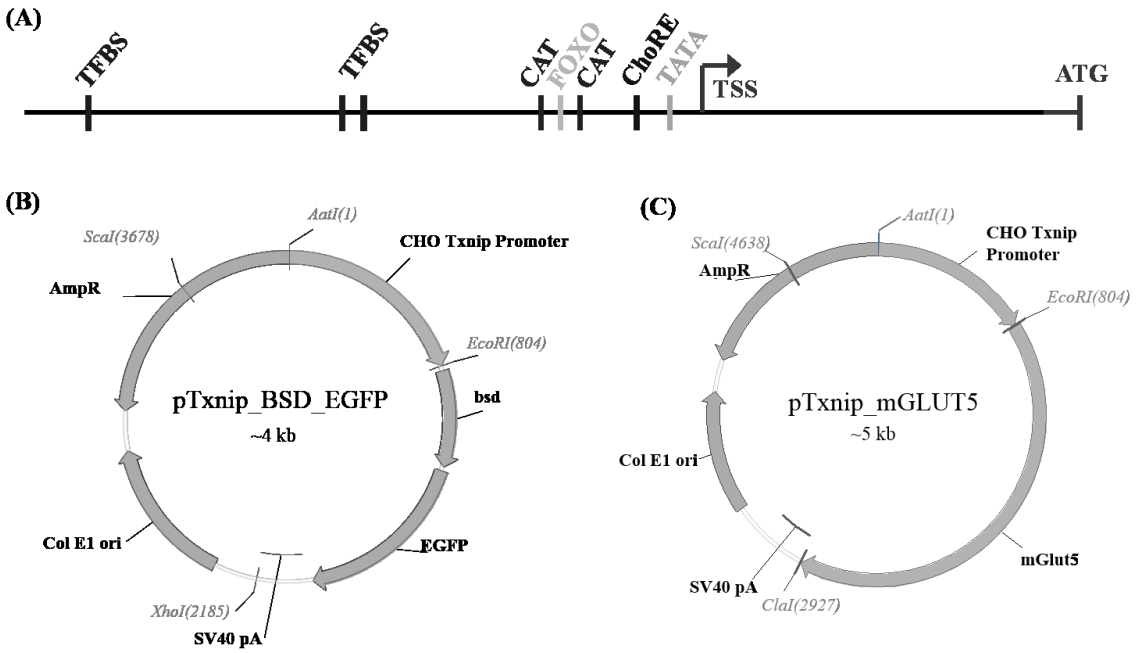


Figure 0-29: Cloning of Txnip promoter from Chinese hamster liver genomic DNA. (A) Isolated fragment of Txnip promoter with approximate locations of the putative transcription start site (TSS), the TATA box, the carbohydrate response element (ChoRE), the CAT boxes, the FOXO binding site, and other transcription factor binding sites (TFBSs). Approximate location of the start codon (ATG) was shown for reference. (B) Map of the expression vector pTxnip_BSD_EGFP. Txnip 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. (C) Map of expression vector pTxnip_mGLUT5. The plasmid includes the mouse GLUT5 gene driven by the Txnip promoter and an SV40 polyA site for transcription termination.

The expected PCR product of ~2100 bp was cloned in the pTxnip_BSD_EGFP plasmid to generate the pTxnip_mGLUT5 plasmid, in which the complete fusion gene (blasticidin S resistance and EGFP genes), was replaced by the mGLUT5 cDNA using EcoRI and ClaI restriction enzymes (Figure 0-29C).

6.4.2 Generation and Characterization of Stable Pools and Clones

All plasmids constructed above were linearized with *ScaI* prior to transfection into the host cells. 24 hr prior to transfection, CHO cells expressing a recombinant protein were seeded in duplicate wells per transfection at 5 x 10^5 cells/mL in 2 mL of DMEM-F12

medium in a 6-well plate. A DNA – cation lipid complex containing 3.6 μg of the linearized pTxnip_mGLUT5 plasmid and 0.4 μg of the linearized pCMV_BSD_EGFP plasmid was used. After incubation for 20 min at room temperature, the complex was added drop by drop into the cells. 6 hr after transfection, the transfected cells were centrifuged at 700 rpm for 5 min at room temperature and re-suspended in 2 mL of fresh DMEM-F12 medium.

Twenty-four hours following transfection, the cells transfected with pTxnip_BSD_EGFP 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 spun down and resuspended in DMEM-F12 medium containing 2 g/L fructose and no glucose 24 hours following transfection. The cells were maintained in this medium for 14 days and then further selected in 10 μg/mL Blasticidin S (Invivogen, San Diego, CA) for 4 days. The selected stable pool (TSGP) was cultured in fed-batch mode using this fructose medium. The expression level of mGLUT5 was characterized over the course of the fed-batch culture using qRT-PCR.

The cloning plates were incubated for approximately two weeks in a 37°C, 5% CO_2 environment. Single clones in 96-well plates were picked and transferred to new wells, each containing 150 μL of fresh selective medium. As cell density increased, the culture was gradually expanded 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.

6.4.3 Fed-batch Cultures

The cultures were initiated by inoculating cells at 0.5 x 10^6 cells/mL, in 20 mL of DMEM-F12 medium in a 125 mL shaker flask (Thermo Scientific, Rochester, NY). The flasks were then incubated at 37°C on a shaker rotating at 130 rpm in a humidified incubator

with 5% CO_2. One mL of a ten-fold concentrated feed medium (10 X) was added daily starting from day 2. In the bi-substrate culture, the cells were given glucose feed medium in the growth stage until day 5 or 6 when the cell reached maximum cell density, after which they were fed fructose feed media to adjust the concentration of fructose to 1 g/L. The bi-substrate cultures with switch back to glucose were fed twice a day to avoid glucose depletion.

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. For hexose measurements, samples were preheated at 65ºC for 20 min. to inactivate phosphoglucoisomerase (PGI) released from dead cells. Glucose was measured using Infinity Glucose Hexokinase Method (Thermo Fisher Scientific, Dubuque, IA), and total hexose was measured similarly with the addition of phosphoglucose isomerase (PGI) (Roche Applied Sciences, Basel, Switzerland) at 35 U/mL and an incubation time of 20 min. Fructose concentration was estimated by subtracting glucose concentration from total hexose concentration. Lactate concentration was measured using the YSI Model 2700 industrial analyzer (Yellow Springs Instruments, Yellow Springs, OH).

6.4.4 Quantitative Real Time Polymerase Chain Reaction (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 listed in Table 0-1, were designed using Primer3 (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 95^0C 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 alongside a no reverse transcriptase control and a no template control. β-actin was used as a reference for comparison across samples.

6.5 Results

6.5.1 Identification and Characterization of Dynamic Promoters

Archived 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. Principal component analysis was used to identify major expression trend followed by k-means clustering to separate the genes into different dynamic categories. Among approximately 1000 genes selected, 15 genes with available 5'-end sequence in our repertoire and showing an upswing dynamic expression pattern were chosen (Figure 0-30).

Among these 15 genes, three genes exhibited upswing dynamics in another cell line. Out of these three genes, the *txnip* gene was selected because its expression level was shown to increase by 23-fold between exponential and stationary phases in another study (Kantardjieff et al, 2010a).

An 800 bp fragment upstream of the translated region of the Txnip gene was isolated from genomic DNA of Chinese hamster liver. The isolated fragment contains approximately 280 bp of the 5'-untranslated region (5'-UTR) and 520 bp of the upstream region of the putative transcription start site (TSS) (Figure 0-29A). The upstream region 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.

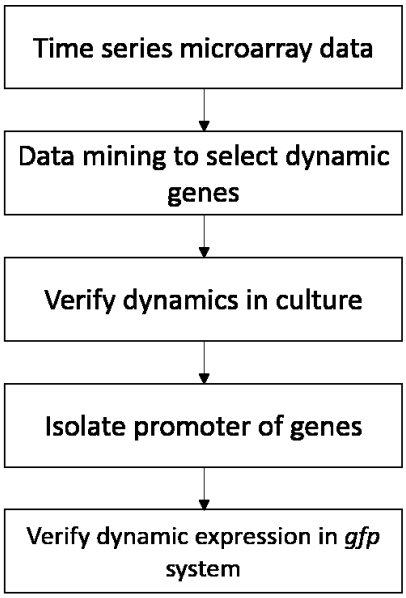


Figure 0-30: Flowchart summarizing the procedure of identify dynamic promoters from time-series microarray data

6.5.2 Engineering dynamic expression of mouse Glut5

CHO cells do not natively express the glut5 transporter. The expression level of the transporter in CHO cells is at least 5-fold lower than the median expression level of 500 from microarray data. The Chinese hamster tissues brain and liver express high levels of the transporter. In contrast the expression level in ovary tissue is also absent similar to CHO cells.

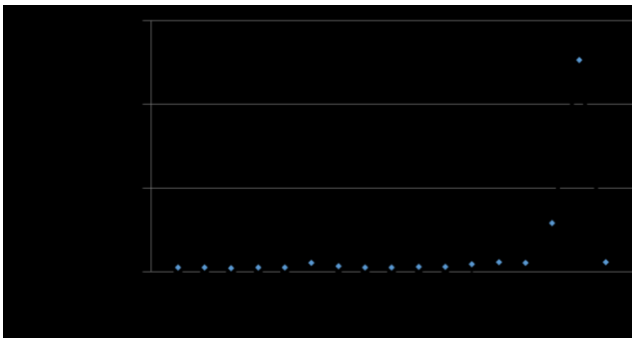


Figure 0-31: Microarray expression level of glut5 transporter (Slc3a5) in several CHO cell lines and Chinese hamster tissue.

In addition to this, the parental cells also failed to grow in completely fructose media. Since the background expression level of the transporter is low, the effects seen in culture can be interpreted as solely resulting from the expression of the transgene.

6.5.2.1 Characterization in mono-substrate culture

The Txnip promoter was further used to drive the expression of a fructose transporter isolated from mouse cDNA, mGLUT5. We have shown previously that CHO cells do not utilize fructose as a nutrient and that the introduction of an exogenous GLUT5 enables them to use fructose (Wlaschin and Hu 2007). CHO cells were transfected with the vector pTxnip_mGLUT5 (Figure 0-29C) and selected in fructose medium followed by selection with blasticidin. Fed-batch culture of the stably transfected pool was performed using either glucose or fructose as the sole carbohydrate source. The cultures were characterized for growth, lactate production, hexose consumption, mGlut5 expression and rIgG production, results are summarized in Figure 0-32. Data from the control glucose culture are shown in open triangles and that from fructose culture are shown in open circles. The cells grow to a higher cell density in fructose culture compared to glucose culture (Figure 0-32A), consistant with what was observed before (Wlaschin and Hu 2007). The glucose consumption rate is high in the growth stages of the culture and steeply reduces in the late stage (Figure 0-32C). The mGLUT5 gene exhibited the dynamic profile as expected, with an approximately 5-fold increase in transcript level on day 10 compared to day 4 as measured by qRT-PCR (Figure 0-32D).

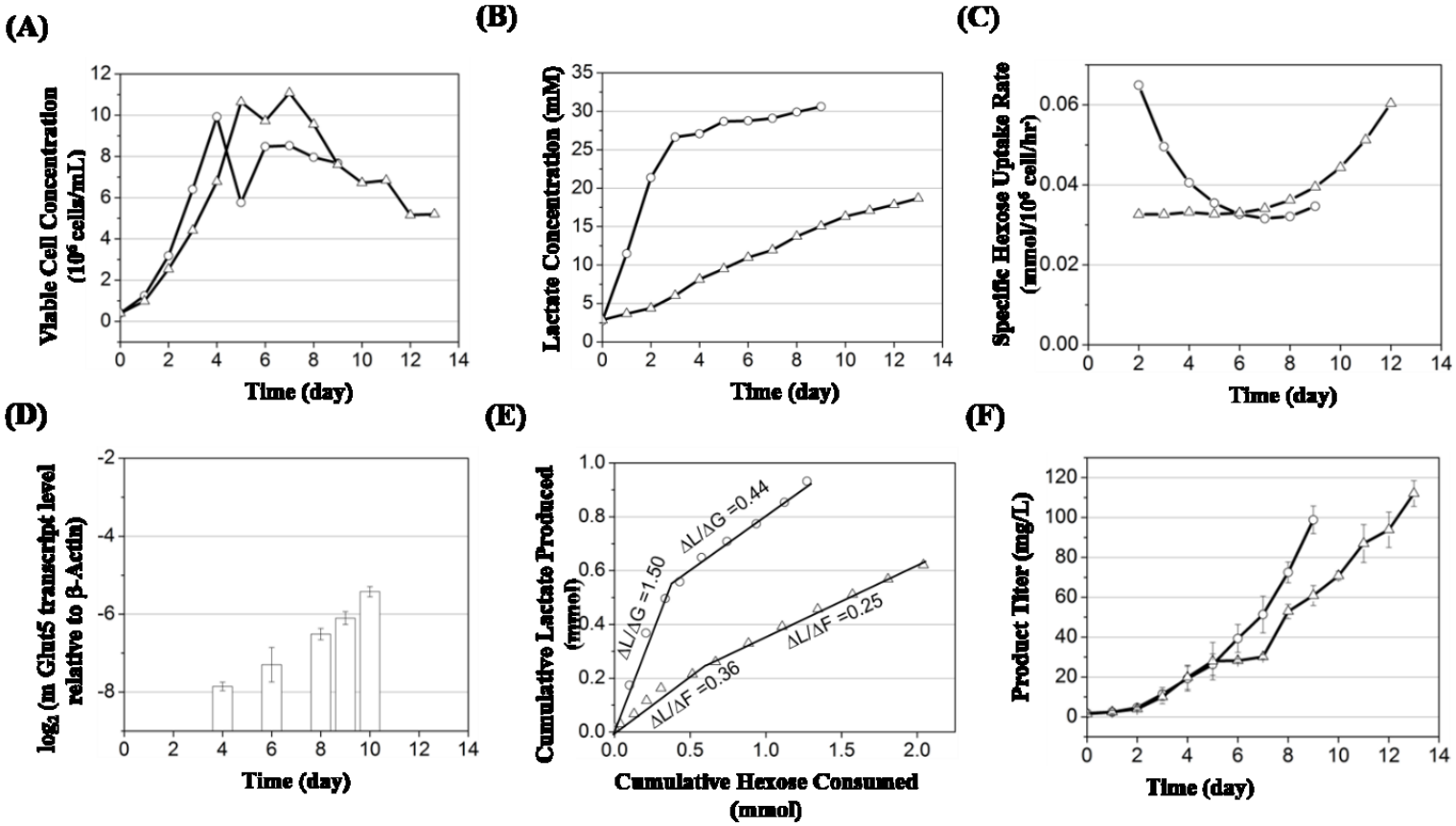


Figure 0-32: Metabolic characterization of TSGP pool stably expressing mGLUT5 gene in a dynamic fashion. Symbols represent data from glucose culture (○) or fructose culture (△). (A) Viable cell concentration of TSGP in glucose and fructose culture showing improved growth and higher maximum cell concentration in fructose culture compared to glucose. (B) Lactate concentration revealing lower rate of lactate accumulation in fructose culture. (C) Specific hexose uptake rate of TSGP. (D) Expression level of mGLUT5 over the course of the fed-batch culture shows dynamic behavior, as well as dynamically increasing specific fructose uptake rate, shown in bar chart. The expression was measured by qRT-PCR. (E) Stoichiometric ratio of mole of lactate produced per mole of glucose or fructose consumed in glucose culture. Cells produced less lactate per mole substrate consumed in fructose culture compared to glucose culture. (F) Titer of TSGP cells in glucose and fructose culture, showing higher titer in fructose culture.

The consumption of fructose and the production of lactate was also measured. The specific fructose consumption rate stayed relatively constant but then increased toward the end of culture period following the increase in the transcript level of mGlut5.

The lactate concentration increased to 30 mM in the glucose culture, whereas the lactate concentration only increased to 19 mM in the fructose culture (Figure 0-32B). The final titer was about 70 mg/L in glucose culture and 119 mg/L in fructose culture, increasing the titer by 40% (Figure 0-32F). Similar to the resulted reported previously, the stoichiometric ratio of lactate to sugar consumption observed was lower than that typically seen when glucose was used as the carbohydrate source (Figure 0-32E). Even though the titer from the glucose culture was higher, the specific productivity is lower (Figure 0-32F).

However, it must be noted that all the above data was obtained from culturing a pool, which is genetically heterogeneous. A few clones of this pool were obtained by dilution cloning. One of the clones was characterized in fed batch culture. These cells also showed growth to higher cell densities in fructose culture and switch to lactate consumption in late stage when low fructose levels were maintained (data not shown).

6.5.2.2 Characterization in mixed substrate culture

To demonstrate the potential utility of dynamic expression, the transfected cells were cultivated in the presence of both glucose (25 mM) and fructose (5mM).

It has been shown previously that at a low expression level of mGLUT5, the consumption rate of fructose is low unless fructose is present at 30 mM or higher (Wlaschin and Hu 2007). Under the culture condition used in this current study, glucose was expected to be consumed quickly by the cells as the main sugar source while fructose would be consumed only at a rather low rate. Such differential consumption between glucose and fructose was seen in Figure 0-33 during the exponential growth stage. Glucose was further added intermittently as shown in Figure 0-33 to sustain a high growth rate.

It has been reported previously that a switch from lactate production in the early stage to lactate consumption in the late stage is a desirable feature in many cultures (Le et al, 2012). Such a switch of metabolism occurs when both sugar consumption rate and

glycolytic flux are low. One way to achieve a low glycolytic flux is to reduce the sugar concentration to such a low level that allows glucose to become almost depleted. However, such a strategy of allowing near depletion of glucose may also trigger apoptosis due to starvation.

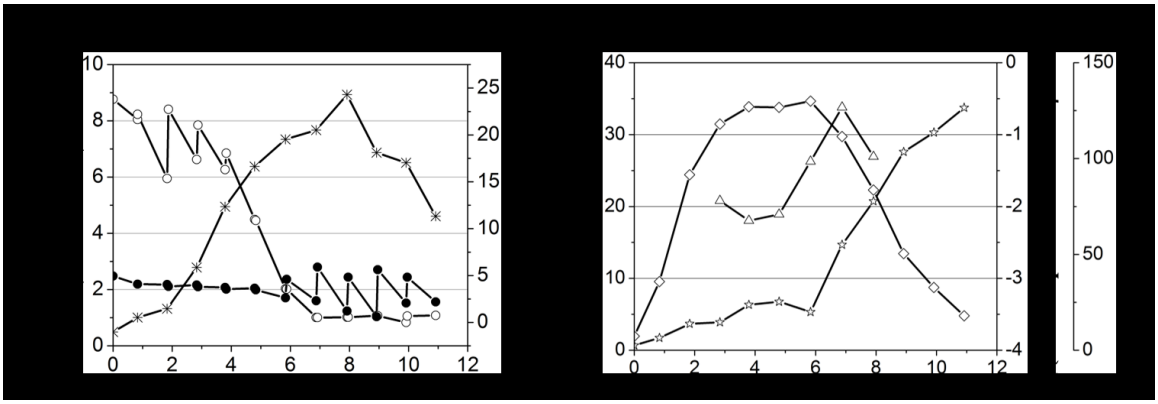


Figure 0-33: Fedbatch culture of transfectants expressing GLUT5 driven by Txnip promoter symbols: concentrations of cells (✳), glucose (○), fructose (●), lactate (◇), and antibody (☆), and relative transcript level of mGLUT5 (△) represented as -log$_2$ (fold change in mGLUT5 expression relative to β-actin).

In this culture, we employed a strategy of switching to fructose consumption at a low rate to avert apoptosis. As can be seen, glucose was allowed to be depleted when the culture reached stationary phase. The increased expression of mGLUT5 in the late stage and the presence of fructose at 5 mM enabled the cells to utilize fructose and sustain glycolysis at a low flux. As anticipated, lactate was consumed along with fructose upon glucose depletion.

6.5.2.3 Investigating the memory effect in metabolic shift

Metabolic shift has been speculated to have a memory effect, i.e. once it is initiated, even if we switch back the sugar to glucose the cells should continue to consume lactate. In order to test this hypothesis, the cells cultured in glucose medium were switch to fructose medium in order to initiate lactate consumption, as shown in the previous section. After a 5-10 mM reduction in lactate concentration, the cells were then transferred back to glucose

feeding at different levels. When the cells were maintained at 3 g/L, 1.5 g/L or 1 g/L of glucose, the cells continued to produce lactate. Only when the cells were maintained at 0.5 g/L, the cells marginally continued to consume lactate (Figure 0-34).

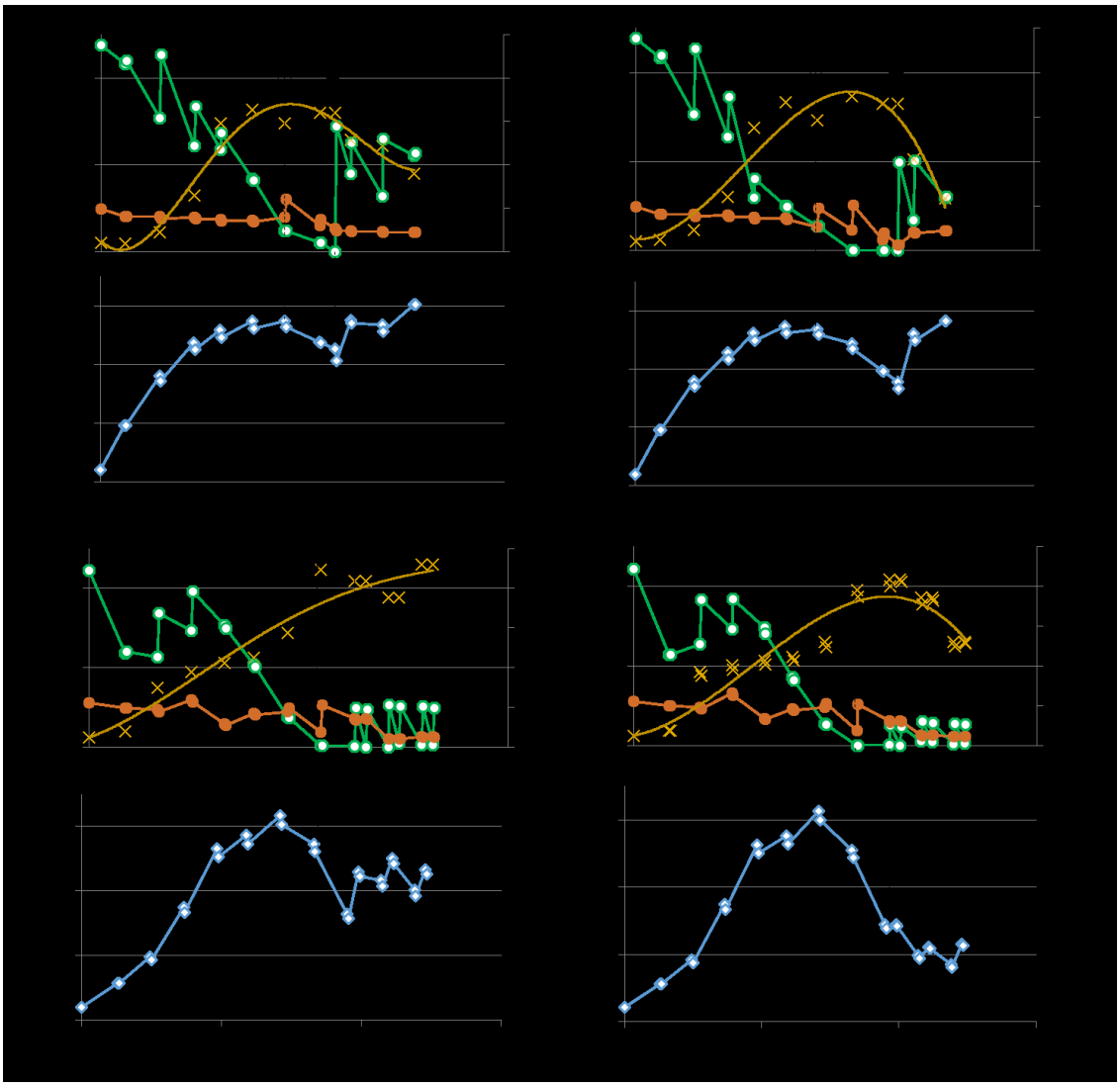


Figure 0-34: Mixed substrate culture with switch back to glucose after lactate consumption. Concentration of cells (×), glucose (○), fructose (●) and lactate (◇) for cells switched to glucose after a 5-10 mM reduction in lactate concentration and then cells were switched to (A) 3 g/L, or (B) 2 g/L (C) 1 g/L (D) 0.5 g/L glucose concentration. Open arrowhead shows time point for switch to fructose and solid arrowhead shows time point at which cells were switched back to glucose.

Also, for maintaining the lower glucose concentrations (1 g/L and 0.5 g/L) the cells had to sampled and fed twice a day to avoid glucose starvation. Despite that effort, glucose was completely depleted by the time the cells were fed, It was also observed that the pH of the culture increased (became more basic) considerably after lactate consumption. The pH was not controlled in the shake flasks.

6.6 Discussion

Time-series transcriptome data to identify dynamic promoters in CHO cells whose expression is dependent on culture stage. We contended that these promoters can be used for the expression of transgenes in cell engineering.

Gene expression in cultured mammalian cells is typically not highly dynamic. For a given gene, the transcript levels in culture usually only changes by a few folds. In the fed-batch cultures investigated in this study, the number of genes whose transcript level changes more than 5-fold on average is fewer than 200. This is in contrast to what is often observed during tissue development *in vivo*, for example in adipogenesis (Soukas et al, 2001); and in stem cell differentiation *in vitro*, where drastic transcription pattern change is not uncommon (Takahashi & Yamanaka, 2006; Ulloa-Montoya et al, 2007).

Based on the identified transcription patterns, 15 dynamic promoters that give similar upswing dynamics under different fed-batch conditions were selected. Among them, three genes (Mmp12, Txnip, and Serpinf1) showed similar profiles in a different producing cell line. Even though both lines were derived from CHO-K1, minute differences in culture conditions and mutations accumulated over long term culturing may have resulted in drastic difference in transcriptional responses.

Since Txnip was also shown to be dynamically expressed in another CHO cell line undergoing temperature shift and sodium butyrate treatment (Kantardjieff et al, 2010a), it was selected for subsequent promoter isolation. A fragment of approximately 800 bp upstream of the putative start codon of Txnip was isolated and used for further analysis. Sequence comparison to the mouse homolog revealed the presence of a TATA box, a

carbohydrate response element (ChoRE), a CAT box, and the binding sites of several transcription factors such as FOXO (Figure 0-29A). The equivalent region in mouse has been shown to harbor a promoter activity at least 5-fold higher than the SV40 promoter (Wang et al, 2005).

It is conceivable 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 profile of Txnip.

The approach taken in this study was through random integration of transgene driven by the promoter. A different approach is to insert the transgene in the locus under the control of the endogenous promoter. Such targeted transgene integration has been greatly facilitated by the advanced in various genome targeting methods, including Zinc finger nuclease (ZFN) (Keeler et al, 1996; Urnov et al, 2005; Urnov et al, 2010), transcriptional activator-like effector nuclease (TALEN) (Boch et al, 2009; Christian et al, 2010; Miller et al, 2011; Mussolino et al, 2011) and clustered regulatory interspaced short palindromic repeats (CRISPR)/Cas-based technologies ((Cong et al, 2013; Mali et al, 2013; Ronda et al, 2014); reviewed in (Gaj et al, 2013)).

The employment of cell's promoter to engineer the dynamic expression of transgenes presents a novel approach to enhance cell's characteristics. Most cell engineering studies use strong, constitutive promoters originated from bacteria and viruses such as CMV or SV40 to drive the transgenes (Seth et al, 2006a; Wurm, 2004). For example, the expression of mGLUT5 gene driven by a constitutive promoter was shown to reduce lactate accumulation (Wlaschin & Hu, 2007a). The expression levels of these genes are almost always high irrespective of the culture conditions or cellular states. Nevertheless, strong and constitutive expression of a transgene is not always ideal; for some transgenes inopportune expression can even be deleterious. The use of an inducible promoter minimizes the unwanted transcription and permits the transgene expression only in the stage during which it is desired (also discussed in (Weber & Fussenegger, 2007). It may also provide a wider range of gene expression. In contrast, for dynamic expression in

sync with cell's cultural rhythm, one may have to select a number of promoters and test a number of clones to achieve the desired dynamic range of gene expression. Several promoter engineering methods have been addressed to achieve high expression ((Brown et al, 2014; Mariati et al, 2014) and reviewed in (Ho & Yang, 2014)), but only a few aim for such dynamic expression. However, the use of an inducible system may not be desirable in large-scale manufacturing due to cost and regulatory considerations. The use of a native dynamic promoter has potential advantages as it provides a more controllable and stable means for modulating the expression of the transgenes with regard to timing and dose.

We employed a pTxnip-driven mGLUT5 to demonstrate the concept. Fructose consumption remained low in the transfected culture in the early stage in spite of the excess of the sugar. The consumption increased only in the late stage of cultivation when the expression of mGLUT5 was high. This surge of fructose uptake led to an increased influx of sugar into glycolysis, however at a lower rate than glucose, and thus averted apoptosis caused by glucose depletion and enabled the culture to switch from lactate production to lactate consumption.

A direct application of this approach in the industry requires other additional characterizations, most importantly that of protein quality. Glycosylation is the most important of the critical quality attributes of a recombinant protein therapeutic (Abu-Absi et al, 2010). The glycosylation precursors UDP-*N*-acetylglucosamine and UDP-*N*-acetylgalactosamine are much lower when cultured in fructose medium compared to glucose medium culture (Ryll et al, 1994), and may potentially affect glycosylation status and hence, the protein quality. However, it is speculated that metabolic shift has a memory effect, i.e., after switching to lactate consumption, the cells may continue to consume lactate even if fructose replaced by glucose. In that case, the cells will be exposed to fructose only for a short time period which may not impact the product quality significantly. Unfortunately, we find that the cells actually switch back from lactate consumption to lactate production stage after replacing fructose to glucose (Figure 0-34). It must be noted that the experiment was not conducted with adequate pH and DO control. When lactate consumption occurs, the pH of the culture media goes down considerably

(more basic). During the switch back to glucose, it is important to maintain more neutral to acidic environment in the culture to enable the lactate symporter (with H^+) on the cell membrane to drive the lactate into the cell.

The concept of dynamic expression in sync with culture stage may also be applied to control apoptosis for process enhancement. Many apoptotic genes are expressed at high levels only when under stress conditions, such as heat shock or addition of histone modification agents, or at the late stage of a fed-batch culture. Over-expression of anti-apoptotic genes has been shown to delay the decline of culture viability. However, over-expression of some anti-apoptotic genes during the exponential phase may impair rapid proliferation, as reported in the case of high-level expression of anti-apoptotic genes retarding cell growth (Dorai et al, 2009; Nivitchanyong et al, 2007). Anti-apoptosis strategies, such as over-expressing anti-apoptotic genes or siRNAs against pro-apoptotic genes may be best carried out using dynamic promoters.

Chapter 7: Developing Genomic Resources for CHO Cells

7.1 Summary

This chapter summarizes the results of the collaborative efforts initiated by our lab to develop and enhance genomic resources for CHO cells. The first part of this chapter describes the assembly and annotation of the Chinese hamster and CHO cell transcriptome. The second part describes our efforts to enhance the Chinese hamster genome assembly. The final part of this chapter discusses the evaluation of two methods for the high-throughput identification of the transgene integration sites in CHO cell lines.

7.2 Context statement

A collaborative effort funded by the Consortium for CHO Systems Biotechnology was initiated for developing and applying genomic resources for Chinese hamster and CHO cell lines by my advisor, Prof. Wei-Shou Hu. Thiruvarangan Ramaraj and Joann Mudge from the National Center for Genome Resources; Faraaz Noor Khan Yusufi from the Bioprocessing Technology Institute; Getiria Onsongo and Kevin Silverstein from the Minnesota Supercomputing Institute at the University of Minnesota; Adam Hauge and Kenneth Beckmann at the University of Minnesota Genomics Center; Mohit Sharma and Prof. George Karypis from University of Minnesota played important roles in the success of this project. Kathryn C. Johnson and Arpan Bandyopadhyay from Prof. Wei-Shou Hu's lab assisted on the transcriptome annotation and the genome quality assessment, respectively. Postdoctoral associates Mingyong Xiong and Hsu-Yuan Fu assisted me in the integration site analysis project.

The transcriptome assembly was done by FNKY using the Trinity assembler, while the genome assembly was done by MS. Gapclosing was done by TR, while NV conducted the re-scaffolding steps. NV and MS conducted quality assessments of the genome and transcriptome. The transcriptome annotation pipeline was automated by GO and executed by NV and KCJ. Genome annotation, genome assembly and the annotation statistics and

figures for transcriptome and genome annotation were generated by NV. GK, WH, KS and JM provided guidance for many of these efforts.

The integration site analysis methods were conceptualized by WH and NV. Initial experiments for integration site analysis were conducted by NV and MX. The pipeline for data processing was developed by NV and executed by HF and MX. The figures for this section were generated by MX, HF and NV.

7.3 Assembly and annotation of the Chinese hamster and CHO cell transcriptome

7.3.1 Introduction

The past few years have seen a major expansion in the genomic resources that are publically available for CHO cells. The Chinese hamster genome and transcriptome have been sequenced, and there is continued interest in studying different genomic aspects of CHO cells using sequencing technologies. While plenty of information is available, it is still not directly usable for omic analyses, mainly because of the lack of useful annotation. Most softwares for functional analysis only accept mouse or human gene symbols. There are also no Chinese hamster-specific ids that can be used for such analyses. For example, the Chinese hamster genome in NCBI has 22,036 CHO transcripts, of which only ~ 30% of annotations are usable for functional analyses. The remaining transcripts have been given 'hypothetical' or 'predicted' gene status. The accurate gene determination is extremely valuable, however, a small dataset of 8,000 genes cannot give us much functional insight from transcriptomic studies.

Our lab initiated a collaborative effort to assemble and annotate the Chinese hamster and CHO cell transcriptome. Historical transcriptome sequencing data in Sanger, 454 and Illumina sequencing platforms were used to assemble the transcriptome using the Trinity assembler (Grabherr et al, 2011). The capability of assembling large datasets while retaining alternative splicing information, made the Trinity assembler suitable for our purposes. A total of 353,445 transcriptome contigs were compiled in the final

transcriptome database. A rigorous annotation pipeline was developed in-house, and used to annotate the assembled transcriptome with homologous gene IDs from related species. The annotated transcriptome assembly has a comprehensive coverage of the CHO transcriptome which is a valuable resource for studying CHO cells.

7.3.2 Results and discussion

7.3.2.1 Transcriptome assembly

Expressed Sequencing Tags (ESTs) with the Sanger sequencing technology were generated by Hu and coworkers in ca. 2005 (Kantardjieff et al, 2009; Wlaschin et al, 2005). An approach of sequencing diverse cell sources, culture conditions, and treatments was used. Several Chinese hamster tissue sources were added to increase the diversity of transcripts. Some of the libraries were normalized to remove redundant copies of highly expressed transcripts.

Illumina sequencing allowed us to sequence to much greater depths and detect rare transcripts. The treatments used to maximize transcript detection included 5-azacytidine (DNA demethylation), tunicamycin (inhibition of N-glycosylation, induction of ER stress, cell cycle arrest), and sodium butyrate (histone deacetylation inhibitor). In total, more than 40 Gbp of transcriptome data was obtained (Table 0-4).

Since the sequencing error profiles in different sequencing platforms vary significantly, the data generated from each sequencing platform was compiled into separate bins. Each bin was assembled individually using the Trinity assembly pipeline (Grabherr et al, 2011). The assembly from the three bins were pooled together as a combined assembly.

Table 0-4: Transcriptome sequencing reads used for the Trinity assembly. Reads from Sanger, 454 and Illumina sequencing technologies were used for the assembly. A total of more than 40 Gbp of transcriptome sequence was used for the assembly.

Description (mRNA Source)	No. of Reads/ESTs	Average Read Length (bp)	Total Output
Sanger Sequencing			
CHO DXB11	105,141	~ 800	84.1 Mbp
Recombinant CHO DXB11 producing IFNγ			
Recombinant CHO DXB11 producing IgG			
Recombinant CHO DG44 producing IgG • Sodium butyrate treated			
Recombinant CHO DG44 producing IgG • 5-Azacytidine treated			
Recombinant CHO DG44 producing IgG • Tunicamycin treated			
Chinese hamster brain			
Chinese hamster spleen			
454 Sequencing			
Recombinant CHO-K1 (IgG) • Normalized library • Exponential growth phase, suspension • GsuI digestion of polyA	399,744	212	84.75 Mbp
Recombinant CHO-K1 (IgG) • Normalized library • Exponential growth phase, suspension GsuI digestion of polyA	262,059	219	57.39 Mbp
Illumina sequencing			
Recombinant DG44 (IgG) • Exponential growth phase, suspension • Serum independent	55,807,972	46	2.57 Gbp
Recombinant DG44 (EPO)- High Producer • Exponential growth phase, suspension • Serum independent • Weak promoter, strong selection	43,766,688	90	3.94 Gbp
Recombinant DG44 (EPO)- Low Producer • Exponential growth phase, suspension • Serum independent • Weak promoter, strong selection	47,611,596	90	4.29 Gbp

Description (mRNA Source)	No. of Reads/ESTs	Average Read Length (bp)	Total Output
Recombinant DG44 (IgG)- High Producer • Exponential growth phase, suspension • Serum independent • 5-azacytidine treatment	49,712,048	90	4.47 Gbp
Recombinant DG44 (GFP)- Unamplified cell line • Suspension culture • 0 mM MTX treatment	6,091,962	90	0.55 Gbp
Recombinant DG44 (GFP)- Unamplified cell line • Suspension culture • 50 mM MTX treatment	6,832,790	90	0.61 Gbp
Parental DG44 • Exponential growth phase, adherent • Serum dependent	36,877,142	90	3.32 Gbp
Chinese Hamster Brain Tissue Brain mRNA from one late adolescent virgin female hamster	47,225,174	90	4.25 Gbp
Chinese Hamster Liver Tissue Liver mRNA from one late adolescent virgin female hamster	73,223,442	90	6.59 Gbp
Parental CHO cell line Untransfected	183,992,242	50	9.20 Gbp

There is a finite possibility that the transcriptome sequences specific to certain tissues are missing in our dataset. In an attempt to include such data, other sources such as, a previous in house assembly (Jacob, 2011) and the NCBI RefSeq database were considered. The sequences from these sources that are unrepresented in our transcriptome assembly were identified using BLAST (Altschul et al, 1990; Mount, 2007). In order to avoid adding spurious transcriptome sequences, a multiple evidence strategy was employed, i.e. only those unrepresented sequences with evidences from both the previous assembly and RefSeq database were added to the combined transcriptome assembly. The remaining sequences were aligned to the Chinese hamster genome to confirm their existence. In this way, 3657 sequences were identified and added to the transcriptome contigs (Figure 0-35). The final number of transcriptome contigs was 353,445.

The assembly statistics for the contigs assembled from each read library are shown in Table 0-5. 248 Mbp of the transcriptome sequence was assembled from Illumina reads alone. While the number of contigs generated was very high, 321,381 of these belonged to 263,967 contig groups. A contig group contains all contigs that are similar in sequence and contain all possible variants of the transcripts. All the variants of the contigs were retained because of the possibility of alternative splicing.

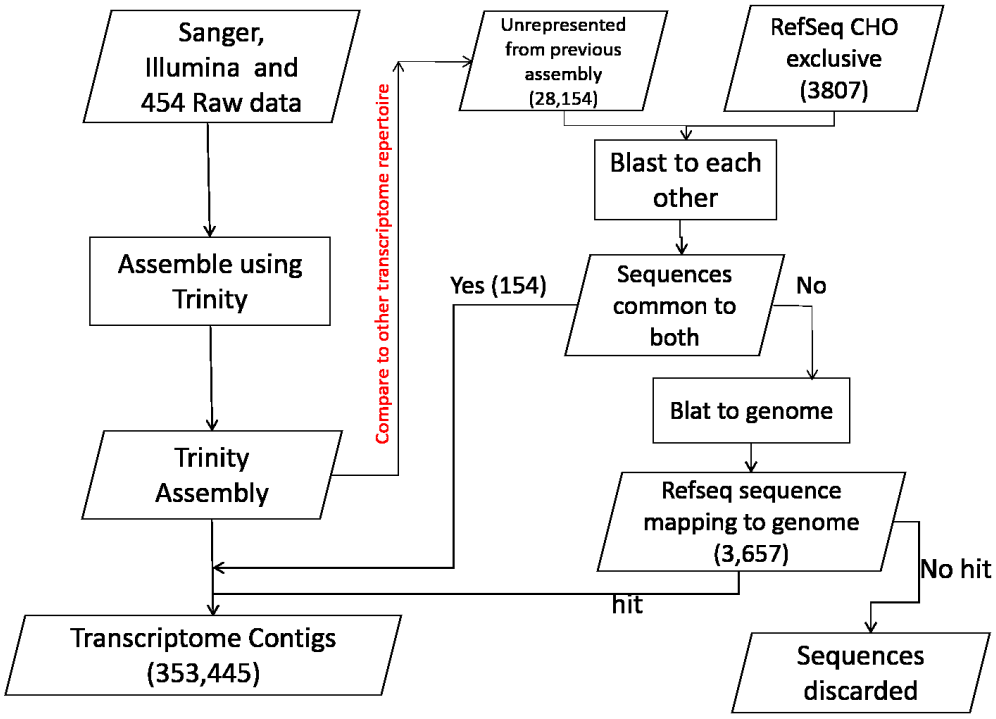


Figure 0-35: Workflow of the transcriptome assembly: The sequencing data obtained from three different sequencing technologies were assembled individually. Transcriptome information from other sources were identified by aligning to the assembly and subsequently included in the transcriptome.

Table 0-5: The number of contigs obtained from each of three sequencing platforms. Majority of the contigs are assembled from Illumina reads. The median contig length of the Illumina contigs is greater than the 454 and Sanger-sourced contigs.

	EST contigs assembled from		
	Sanger Reads	**454 Reads**	**Illumina Reads**
Number of contigs[a]	9,720	18,533	321,381
Number of contig groups[b]	9,418	17,239	263,967
Total sequence (Mb)	8.12	6.70	248
Min contig length (bp)	201	201	201
N50 contig length (bp)	889	369	1,597

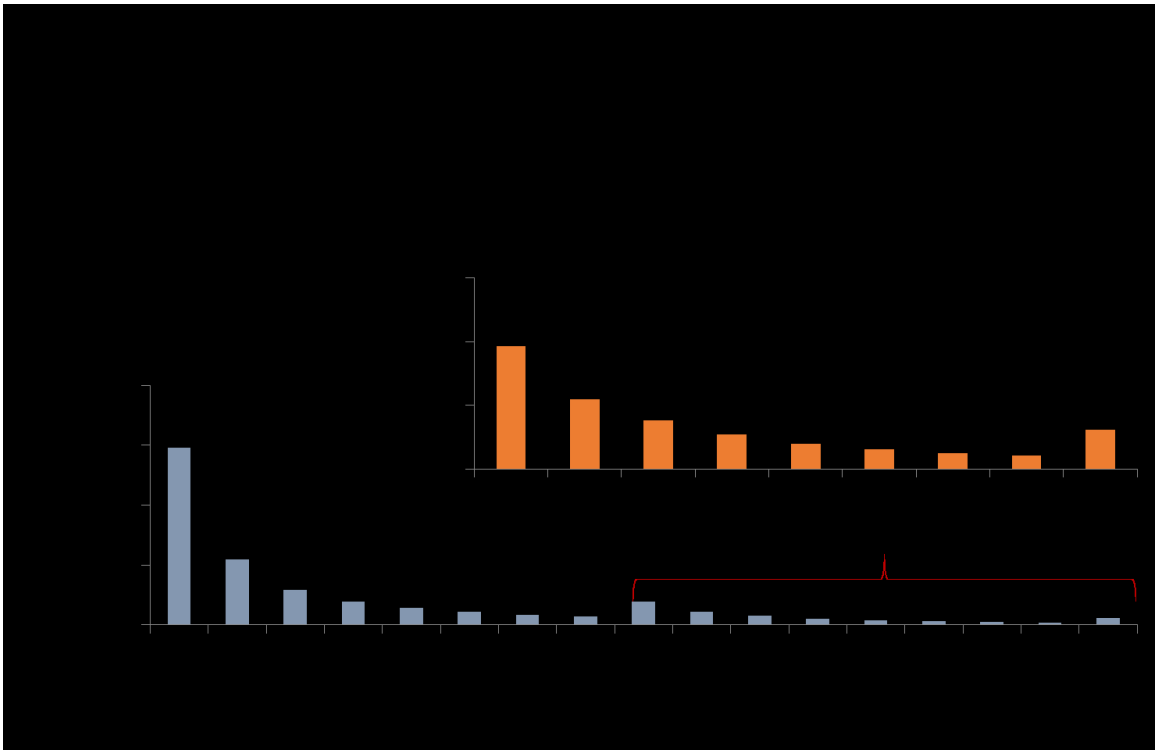


Figure 0-36: Contig length statistics of the final combined transcriptome contigs: Table shows basic statistics. The bar chart shows the contig length distribution. Bar chart in the inset shows contig length distribution for contigs greater than 1000 bp in length.

The assembly statistics of the transcriptome shows an average contig length of 769 bp and an N50 contig length of 1563 bp. N50 is an estimate of the median contig length. It is calculated by sorting the contig lengths in descending order followed by calculating the minimum contig length representing half of the total assembly. The estimate of average mRNA length in humans is approximated to be 1400 bp. Hence, most of the genes represented in the assembly can be assumed to be near full-length. The chart illustrating the size distribution of the EST contigs shows that most of the contigs are small in length. However, the number of contigs more than 1,500 bp is significantly large (Figure 0-36).

Even though the transcriptome contigs are not near full length, the number of contigs is still high enough to be alarming. The large amount (40 Gbp) of sequence information can potentially lead to the assembly of many variants that may have very low representation in the sequencing data. Since the sequencing was conducted in 2009 when the Illumina sequencing technology was not as mature as it is today, these sequences could have also arisen from reads that may have suffered PCR or ligation errors.

7.3.2.2 Transcriptome annotation

The transcriptome contigs were annotated using an in-house homology-based annotation scheme. The schema is explained in Figure 0-37. All the transcript contigs were repeat masked to remove sequences related to the known repetitive DNA elements. These repeat masked contigs were aligned using BLAST to the databases of annotated transcriptome sequences from other species, including Ensembl mouse, Fantom 3, Ensembl rat, Ensembl human, Noncode v3 ("a" in Figure 0-4). The annotation from the alignment is then assigned a tier based on the strength of the BLAST score. A higher BLAST score will result in the annotation being categorized within a high tier.

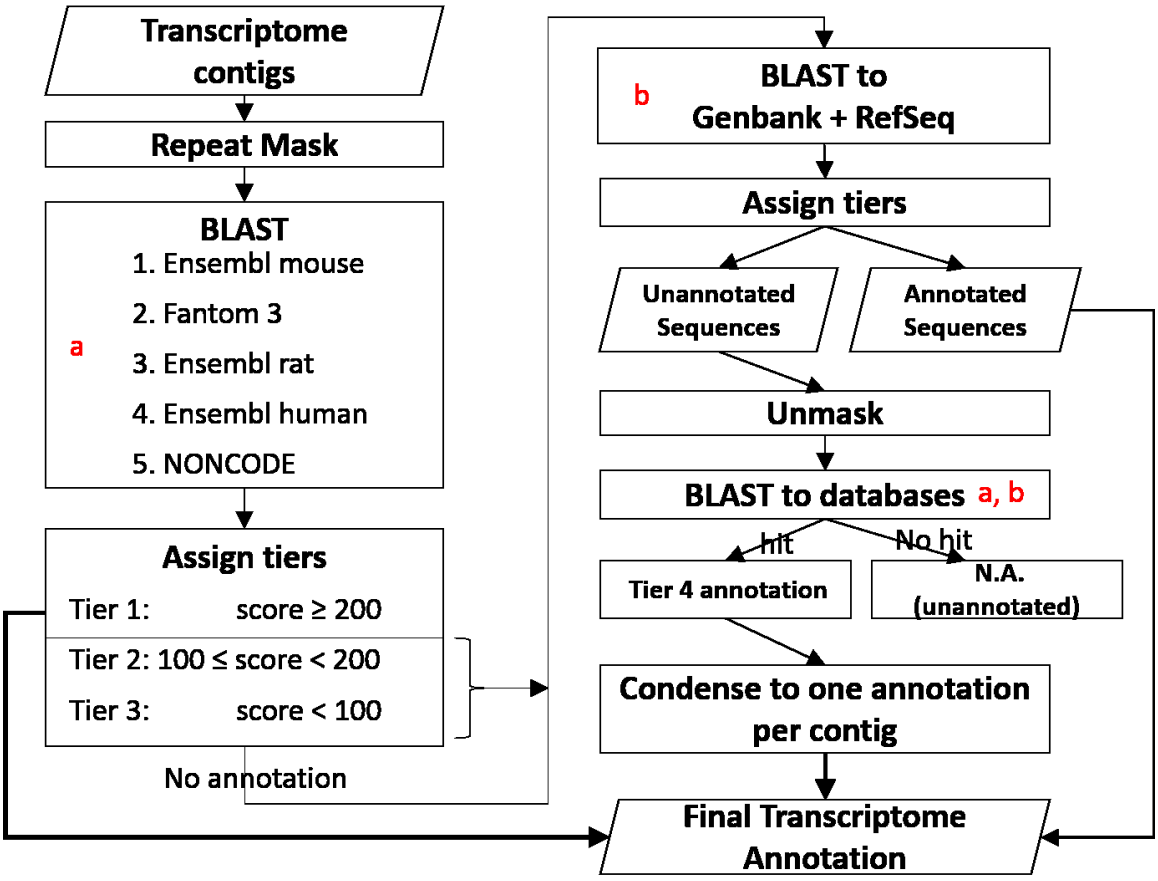


Figure 0-37: Detailed homology-based transcriptome annotation pipeline. Firstly, all the transcriptome contigs were repeat-masked to conceal the repetitive sequences. This repeat-masked sequence was then aligned to the complete dataset of transcriptome sequences from mouse, rat and human sources collected from Ensembl and Fantom databases. The sequences were also aligned to NONCODE database to annotate noncoding RNA. The blast hit scores were used to categorize the annotation by strength into three tiers. Contigs with the strongest tier 1 annotation were directly added to the final transcriptome annotation, and the annotation from the database showing the strongest hit was transferred to the contig. If the contigs hit to all the databases with almost equal scores, then the annotation was given according to a pre-decided database priority shown in the figure. Tier 2, tier 3 and unannotated sequences were aligned to further lower priority databases- Genbank and RefSeq. If significantly better hits are obtained, then the respective annotation is transferred; otherwise the original annotation is retained. Based on the BLAST score, the annotation was given tier 2 or tier 3 status. The remaining unannotated sequences are unmasked and subjected to the same annotation pipeline. The final transcriptome annotation includes one annotation per contig along with a tier assignment.

A contig with blast score higher than 200 to a certain reference transcript was given a tier 1 status. A score between 100 and 200 was categorized as tier 2, and a score less than 100 was assigned a tier 3 status. Within each tier, a higher priority was given to the annotation arising from Ensembl mouse. The priority list of databases is shown in Figure 0-37. If the annotation score from a lower priority database was higher than that from a higher priority database, then the annotation from the lower priority database was assigned only if the difference between the scores is greater than 20.

Tier 2 and tier 3 contigs along with the unannotated contigs were then aligned to the lowest priority database "b", which contain sequences from GenBank (Benson et al, 2014). The GenBank database is a publicly available database of sequence information deposited by individual laboratories. The sts, gss, refseq_rna, nt, est and patnt databases were used. A similar strategy was used to assign tiers to the annotations. Similar to the previously described case, a GenBank annotation was assigned only if the BLAST score to the GenBank sequence was significantly higher than the BLAST score to the databases in the group "a".

The sequences that remain unannotated were then unmasked, and aligned to all the databases. The annotation with the highest score was recorded for each of these contigs.

Table 0-6: Distribution of contigs among the four annotation tiers. Most contigs have tier 3 annotation. Only a few contigs remain unannotated.

Tier	Number of contigs
Tier 1	94,282
Tier 2	84,422
Tier 3	165,242
Tier 4	9,321
No Annotation	178
Total	353,445

The distribution of the contigs between the tiers is given in Table 0-6. More than 25% of the contigs have tier 1 annotation. Majority of the contigs have tier 3 annotation.

Only 178 contigs remained unannotated following the complete execution of the annotation pipeline.

Most functional class analysis software use curated gene sets, pathways or interaction networks, which are only available for human or mouse genes. This can be attributed to their popularity in genomic research, and the fact that most of the interaction data has been generated for human and mouse genes. Often functional class analysis in CHO cells is conducted using mouse gene ids because it is closest relative to Chinese hamster among the rodents. Furthermore, we can avail of the vast amount of functional interactions and pathway information available for mouse. It is for this purpose that Ensembl mouse database was given the highest priority in the annotation pipeline.

Table 0-7: Breakdown by database within the tier 1 contigs. Ensembl mouse was given the highest priority in the categorization, therefore most contigs have Ensembl mouse annotation.

Database	Number of sequences	Number of unique Gene Ids
Ensembl mouse	84,906	18,416
Fantom3	4,168	2,725
Ensembl rat	1,816	932
Ensembl human	1,515	882
NONCODEv3	1,254	459

The distribution of the assigned annotations among the databases within the tier1 contigs shows that close to 90% of the sequences have mouse annotation, representing 18,416 mouse genes (Table 0-7). Second priority is given to Fantom3 which is a manually curated database of functional annotation in mouse generated by the RIKEN consortium (Carninci et al, 2005). Only a few tier 1 contig sequences were assigned to annotation from Ensembl rat and human and NONCODE databases.

Among the sequences in all the tiers, the most represented database is still Ensembl mouse, however, only ~60% of the sequences have Ensembl mouse annotation (Figure 0-38A). The average nucleotide identity between mouse and Chinese hamster is 95% (Figure 0-38B). 24,652 Ensembl mouse genes were covered out of a total of 39,179 total mouse genes.

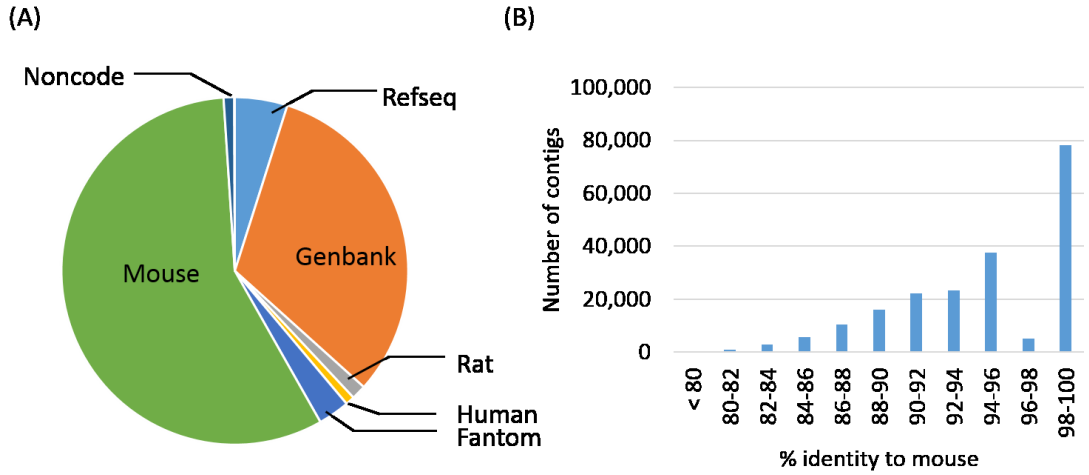


Figure 0-38: (A) Distribution of annotation by database for all the contigs. Most contigs have been annotated to mouse. Only a few contigs show better hits to human and rat. Many contigs also have annotation to Genbank. (B) Distribution of percent identity to mouse transcripts. Most transcriptome contigs have high identity (95-100%) to mouse.

Most of the functional analysis tools also require that all the annotations are of the same type. For example, either all genes in mouse should be represented entirely by their gene symbols, or entirely by MGI ids or by Ensembl mouse gene ids. Therefore, all the annotations need to be condensed to one type of id. Ensembl mouse id was selected because linking information is more easily available from Ensembl ids to other databases.

All the annotations from Ensembl human and rat databases were converted to the corresponding mouse annotations using orthology information. Further, Fantom3 gene ids were also converted to the corresponding Ensembl mouse id. Out of 39,179 total mouse genes, 24,932 mouse genes were finally covered in our database. The mouse gene database also includes pseudogenes that may or may not have an evidence of expression, as well as genes that are expressed in specific tissues that we do not have in our database.

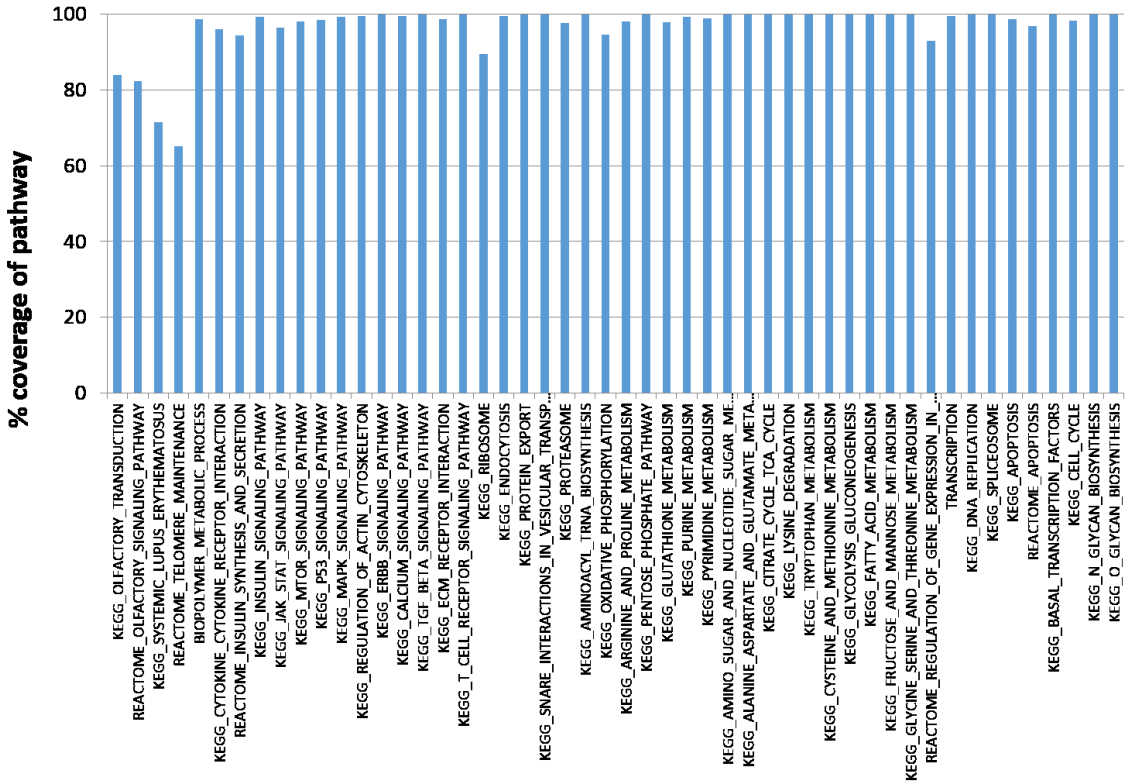


Figure 0-39: Coverage of key pathways: The gene sets for some pathways curated by KEGG and REACTOME were collected for mouse. The corresponding mouse ids in the annotated transcriptome were used to estimate the coverage of the pathway. Most pathways relevant to bioprocessing were well covered in the transcriptome assembly.

Exploring the coverage of major functional groups that are relevant for bioprocessing, we find that most functional classes are completely covered (Figure 0-39). The gene sets were obtained from KEGG pathway database (Kanehisa et al, 2014). The functional classes with lower coverage 60-80% are from pathways such as olfactory transduction, olfactory signaling, systemic lupus erythematosus and telomere maintenance. A few genes in these pathways are probably tissue specific or disease related, because of which they were not represented in our transcriptome assembly.

7.3.3 Conclusions and future outlook

Resource poor species like Chinese hamster can greatly benefit from next-generation sequencing. A large amount of transcriptome sequence data from diverse sources was used to assemble 353,445 transcriptome contigs. A major disadvantage of the high throughput sequencing approach is the large number of contig sequences generated by the assemblers. Even though they represent only a few genes, an ideal algorithm should be able to combine these sequences into longer and fewer contigs.

Although sequencing and assembly methods are available in plentitude, the annotation of the assembled genome or transcriptome is a unique challenge. Annotation strategies are almost species-specific, so most of the tools have to be derived in-house. The annotation pipeline developed for annotating Chinese hamster transcriptome can be easily extended to other resource-poor species with a few alterations. The automated feature of this pipeline enables easy updates as more data becomes available for the reference genomes.

A good quality reference transcriptome assembly will prove to be valuable for using omic approaches to understand and improve the role of CHO cells as recombinant protein producers.

7.4 Assembly and annotation of Chinese hamster genome

7.4.1 Introduction

Transcriptomic studies can reveal immense information about the molecular changes occurring in a given process. Information can be gathered about the transformation of CHO cells from non-producers to hyper-producers with productivities equivalent to nature's professional secretor cells like plasma cells or liver cells. However, the most common form of intervening cells is still manipulating the genome, usually to permanently alter the transcription of some target genes.

In the context of recombinant protein production, it is desired to target the recombinant gene to a 'hotspot' to ensure high and stable transcription of the recombinant gene of interest. Efficient tools like CRISPR/Cas9, Zinc Finger Nuclease (ZFN) and Transcription Activator-Like Effector Nuclease (TALEN) have made targeted genome engineering accessible for use in bioprocessing. For these technologies to be applied for bioprocessing, a good reference genome must be available. It is also of interest to determine the integration site of a cell line to ascertain its propensity to become unstable over the course of expansion to the production stage. Such investigations necessitate having access to the organism's genome.

The previous section described the challenges in transcriptome assembly and annotation. One of the major challenges was the large number of contigs generated by the assemblers for high throughput sequencing data. The genome can be used as an aid to put together contigs from the same gene into full length transcripts.

To this end, the CHO-K1 genome was sequenced in 2011 through a collaborative effort (Xu et al, 2011). However, knowing that CHO cells are genetically heterogeneous with many chromosomal aberrations (Worton et al, 1977) and aneuploidy (Kurano et al, 1990), a reference sequence that is more invariable is suitable as a standard for genomic studies. So, our group initiated an effort to sequence, assemble and annotate the Chinese hamster genome in early 2010. Midway into our efforts, an independent group published the sequence of the Chinese hamster genome (Brinkrolf et al, 2013; Lewis et al, 2013).

This opened up an opportunity to improve the quality of the Chinese hamster genome assembly.

Current technologies for assembly cannot be easily modified to accommodate reassembly or assembly improvements. The algorithms are either too memory intensive or take a long processing time. We used a two prong approach to improve the genome assembly: (1) using an in house assembly as the reference, and information from the public Chinese hamster sequencing to improve the assembly, and vice versa; (2) using the public assembly as the reference and the information from our in house sequencing to improve it. The decreasing cost of sequencing has fuelled a lot of efforts to sequence genomes, and such events requiring the combination of different genome assemblies or improving assemblies will be encountered often in the near future. Our approach and learning will definitely help similar efforts in other organisms.

7.4.2 Materials and methods

7.4.2.1 Source of genomic DNA

DNA was extracted from the liver of a single, highly inbred, female Chinese hamster of the 17A/GY strain (Cytogen Hamsters, West Roxbury, MA). DNA was isolated using the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA) using standard manufacturer recommended protocols.

7.4.2.2 DNA library preparation and sequencing

The extracted DNA was fragmented by nebulization and ligated with Illumina sequencing adaptors to facilitate sequencing (Illumina, CA). For the short insert libraries, the DNA was size-selected using agarose gel electrophoresis. Fragments with insert sizes of 300 bp, 400 bp and 500 bp were selected and gel-purified. The size distribution of the DNA fragments was verified in a 2100 Bioanalyzeer using a DNA 7500 chip (Agilent, Santa Clara, CA).

For the long insert libraries, the DNA was fragmented, end labelled with biotin, and then size selected for 3kb. The DNA fragments were circularized by ligation of either ends

of the linear fragment. The DNA was then fragmented and then the portion of the DNA near the site of ligation was selected using streptavidin bound magnetic beads. Illumina sequencing adaptors were ligated to these DNA fragments and then size selected prior to sequencing on an Illumina GAIIx flow cell. DNA-PET was used to generate 10 kbp and 20 kbp long insert mate-pair libraries. Illumina v3 chemistry was used and the base calling was done by the GA 1.4.2 pipeline using recommended standard parameters.

7.4.3 Results and discussion

CHO cells, since their isolation in 1958 (Tijo & Puck, 1958) have been subjected to many mutagenic agents, many transfections and antibiotic selections. They have also been maintained in culture in many laboratories for many years. Accumulation of mutations over culture time is quite common along with chromosomal abnormalities and aneuploidy. Even though, the Chinese hamster has 11 chromosome pairs, CHO cells usually have a modal chromosome number of 20 or 21 (Worton et al, 1977; Wurm & Hacker, 2011). Several CHO cell lines derived from the mutagenesis of the first CHO cell derived by Puck also have several differences in chromosome number (Derouazi et al, 2006).

The chromosomal variability of CHO cells in culture prompted us to choose Chinese hamster as a reference genome instead of CHO cells. The 17A/GY strain of a female Chinese hamster was sequenced.

7.4.3.1 Sequencing of the Chinese hamster genome

The whole genome was sequenced by high throughput Illumina sequencing. In Illumina sequencing, the DNA is at first fragmented and then size selected to allow a narrow size distribution. Following this, sequencing adaptors are ligated to both ends of each fragment. Sequencing is done from both of the ends to generate a pair of reads for each fragment. The distance between the two reads is known *a priori* from the size distribution. Insert sizes of 300 bp, 400 bp and 500 bp were used in sequencing. The distribution of sequencing data among the libraries is shown in Table 0-8. A total of 273

Gbp of sequence data was used to assemble the Chinese hamster genome. Considering a genome size of 2.66 Gbp (Greilhuber et al, 1983), this amounts to about 100 X coverage of the genome.

Table 0-8: Genome sequencing data- Short insert (300, 400 and 500 bp) and long insert sequence reads (3000, 10000 and 20000 bp) were obtained. A total of 273 Gbp of genome sequence was used for the assembly.

Library type	Number of Reads ($x10^{-6}$)	Insert Size (bp)	Read Length (bp)	Total Output (Gbp)
Paired End	796	300	90	71.65
Paired End	513	400	150	76.93
Paired End	499	500	150	74.89
Mate Paired End	235	3000	54	12.68
DNA-pet	272	10000	76	20.71
DNA-pet	216	20000	76	16.44
Total	**2766**	**------**	**------**	**273**

7.4.3.2 Data pre-processing and assembly

The sequencing data was pre-processed to remove low quality reads, reads with low complexity and trim the adapter sequences. Any read that has a consecutive 15bp match with a primer sequence was rejected along with its pair. A minimum quality cut-off of 5 on a scale of 0-40 was set for each base to be accepted. The first 3 bases of a read and the end of a read are trimmed off if they fall below this threshold. After trimming, only sequences longer than 32 bp were retained. Low complexity filtering was carried out to further improve the read quality. Any read containing 20 consecutive bases of low complexity was flagged. If the remaining non-low complexity parts of the read added up to 32 bases, then the read was kept.

High throughput sequencing reads are usually short sequences that have to be assembled into contigs based on evidence of overlaps. Short read assemblers like AbySS

(Simpson et al, 2009) or SOAPdenovo (Luo et al, 2012) and ALLPATHS (Gnerre et al, 2011) are designed to accommodate such short reads for assembly. Based on critical evaluations from (Bradnam et al, 2013; Earl et al, 2011; Salzberg et al, 2012), we decided to use the AbySS assembler for assembling the Chinese hamster genome.

Sometimes for genome assemblies from high throughput sequencing, the sequencing data is preprocessed to correct for sequencing errors. Since high throughput sequencing generates gigabytes of data in a single run, even a small error rate results in a large number of erroneous base calls that is spread out among the sequencing reads. Some assemblers include an error correction module to correct such reads. Error correctors scan the reads for k-mers with low frequency of representation and substitute a few bases to increase its representation frequency. This is done under the premise that k-mers with much lower frequency than the median frequency probably arise from sequencing errors. In some cases, it is found that the resulting assemblies show better contiguity. In order to explore this possibility, we conducted error correction using the Musket substitution error based corrector, which is the major source of error in Illumina sequencing (Liu et al, 2013).

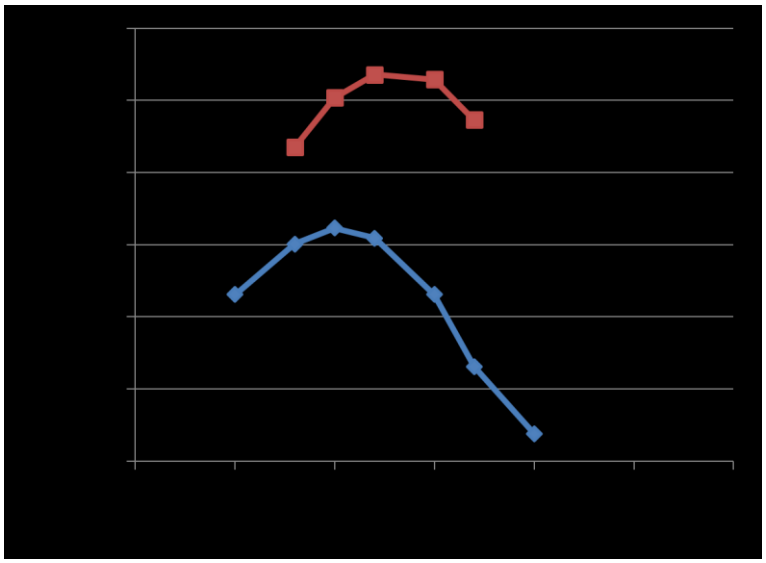


Figure 0-40: k-mer sweep for in house assembly of the Chinese hamster genome. Several k-mer sizes from 65 – 80 bp were assessed for the transcriptome assembly. Prior to error correction (◆), the optimum k-mer was 70 bp, and for the error corrected reads (■), the optimum k-mer increased to 72 bp.

AbySS is a typical de Bruijn graph-based genome assembler, where the reads are broken into k-size fragments, and then connected to each other based on evidence from individual reads. The paths with a high degree of confidence were joined together to make up the genome contigs. The parameter 'k' is very important for the optimal performance of the assembler, and has to be optimized before next step. A series of k-mer values from 65 – 80 bp were tested on the short insert paired end reads in the first step of the assembly. The optimum parameter was found to be 70 for the preprocessed reads, and 71 for the preprocessed and error corrected assembly.

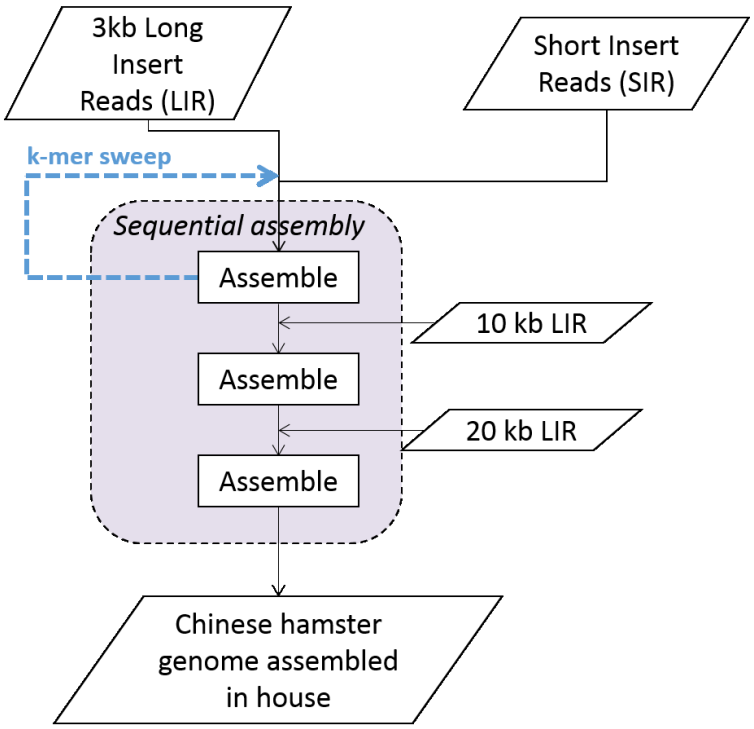


Figure 0-41: Sequential assembly workflow for Chinese hamster draft genome – Following k-mer optimization, the optimum k-mer was used to assemble the short insert reads into contigs using AbySS. The 3 Kbp long insert reads were used to scaffold together the contigs from the assembly. Subsequently, the 10 Kbp long insert reads were used to re-scaffold the scaffolds from the previous stage of assembly. Similary, the 20 Kbp long insert reads were then used for further scaffolding. This genome assembly is referred to as the in-house assembly.

We chose to use the assembly that was done without the error corrected reads because the improvement after error correction was not very significant (Figure 7-6). The

contig N50 at optimum k-mer was 4.2 Kbp for preprocessed reads, and improved only to 4.4 Kbp for the error corrected assembly.

The next step after contigging is scaffolding, which uses the long insert reads for establishing long-range ordering of the contigs, even if the intervening sequences between the contigs is unknown. Long-insert reads are essential for bridging repetitive and/or low complexity sequences that are longer than the short-insert size. The final scaffold is a series of ordered contigs separated by stretches of N's, where the length of the stretch is an estimate of the distance separating the contigs on the genome. The scaffolder estimates the insert size of the library by mapping the reads to the contig assembly. Hence, in order to be able to bridge the contigs using long insert reads, the contig length should be greater than the lowest insert size of the libraries used for scaffolding, or preferably longer than most of them. Even though the lowest insert size was 3 Kbp, the N50 of the contig assembly was 4.2 Kbp. However, it is still lower than the physical length of 10 Kbp and 20 Kbp insert size libraries. The standard scaffolding procedure had to be modified to overcome this limitation. A sequential assembly strategy was employed. The 3 Kbp long insert libraries were used to conduct the preliminary assembly. The 10 Kbp long insert reads were used to scaffold the preliminary assembly, and create a secondary assembly. The secondary assembly was then scaffolded using the 20 Kbp long insert reads to obtain the final Chinese hamster in-house draft assembly (Figure 0-41).

7.4.3.3 Integration of public genome data into Chinese hamster genome assembly

Two independent efforts towards the sequencing and assembly of the Chinese hamster genome were published recently. Some of the sequencing data was made publically available (Brinkrolf et al, 2013; Lewis et al, 2013). We used this opportunity to use the publically available data from these projects to enhance our in-house assembly. Only a portion of the raw sequence data was made publically available, so conducting a combined assembly was not possible. We had to develop a unique approach to integrate the information from public sources into our in-house assembly. We used the following

strategies- (1) Using the in-house assembly as the reference, we used the limited public sequencing data to enhance its contiguity (Figure 0-42A) (2) Using the raw sequencing reads from in house efforts, we attempted to improve the public Chinese hamster assembly (Figure 0-42B).

The in-house assembly had 19% gaps in the genome (Table 0-9), a relatively high fraction for a genome assembly. This was much higher than the 2.5% gaps in the public assembly. So we chose to employ a gap filling approach to improve in-house data. By mapping the short insert reads back to the assembly, it can be locally reassembled to improve the contiguity of the draft genome. Gapcloser was used for this purpose (Tsai et al, 2010). The short insert reads (SIR) from public repository and the short insert and long insert reads (LIR) from our in-house sequencing efforts were used to fill in the gaps in the in house genome assembly (Figure 0-42A). By this approach, the gaps reduced from 19% to 8% (Table 0-9). Because of the high fraction of gaps, the software ran for 70 days until completion.

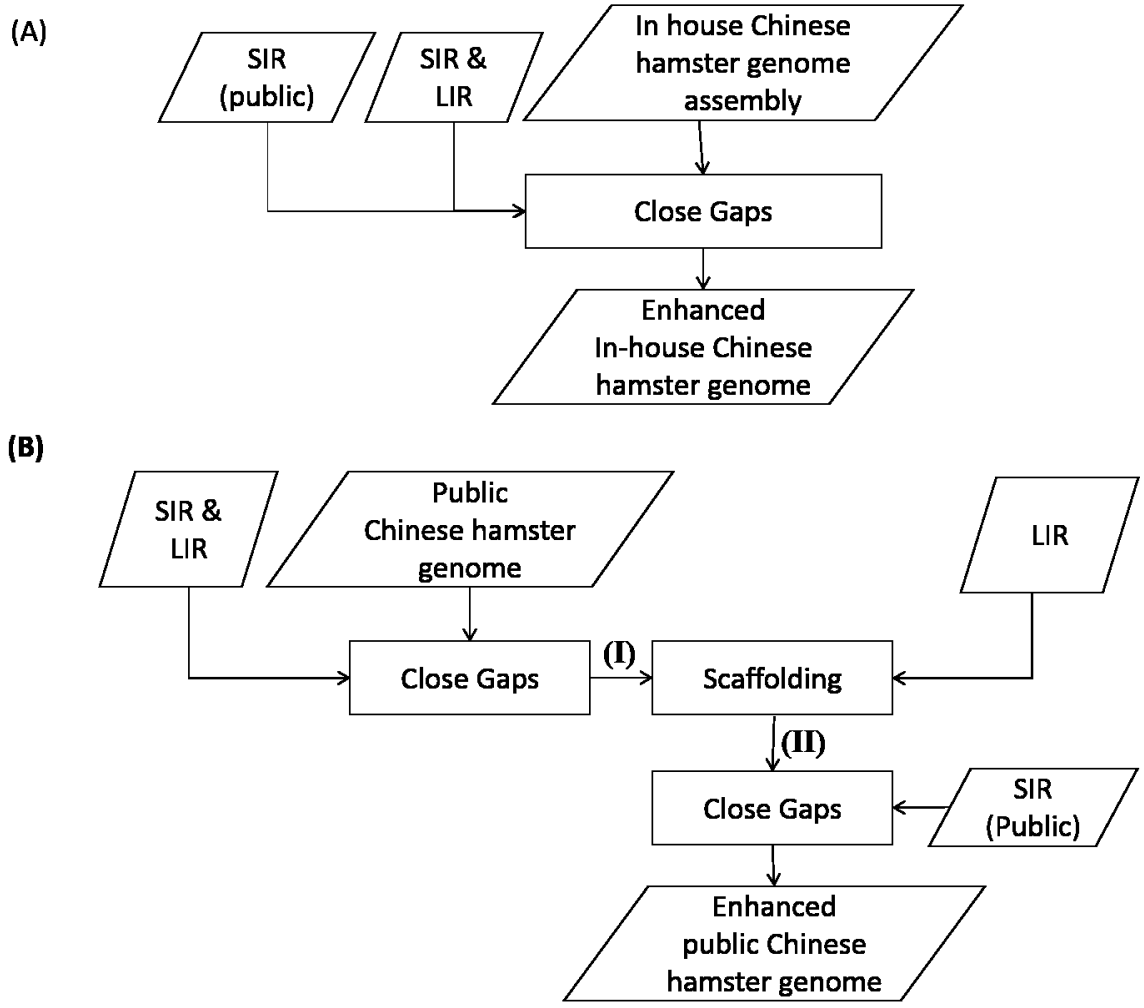


Figure 0-42: Workflow for genome assembly improvement. (A) Short insert reads (SIR) from public genome sequencing efforts and short and long insert reads from in house efforts were used to close the gaps in the in house assembly. This is the final enhanced in-house assembly. (B) In-house generated SIR and long insert reads (LIR) were used to close gaps in the public Chinese hamster genome draft assembly. This assembly was re-scaffolded using the long-insert reads, followed by further gap filling using public SIRs.

Table 0-9: Draft genome statistics at different stages of genome assembly enhancement showing continual improvement in contiguity after each step.

	Enhancement of the in house genome		Enhancement of the public Chinese hamster genome			
Parameter	**In house genome**	**Enhanced in house genome**	**Public genome**	(I)	(II)	**Enhanced public genome**
Contigs	694,058	314,050	176,068	46,025	46,025	87,104
Max Contig (bp)	55,898	109,870	219,443	804,483	804,483	961,385
Mean Contig (bp)	2,933	7,376	12,980	50,189	50,189	26,757
Contig N50 (bp)	4,036	13,950	27,317	127,489	127,489	121,600
Contig N90 (bp)	1,458	3,363	7,453	34,453	34,453	32,129
Total Contig Length (bp)	2,035,409,935	2,316,301,122	2,285,446,913	2,309,963,092	2,309,963,092	2,330,639,248
Assembly GC (%)	41.69	41.43	41.39	41.42	41.42	41.43
Scaffolds	83,691	83,691	10,868	10,867	7,876	49,719
Max Scaffold (bp)	16,416,848	16,386,994	8,324,132	8,320,503	60,518,007	60,517,459
Mean Scaffold (bp)	30,143	30,080	215,691	215,652	298,330	47,579
Scaffold N50 (bp)	2,593,447	2,580,348	1,579,055	1,578,388	17,211,978	17,210,817
Scaffold N90 (bp)	182,764	180,187	415,923	415,119	3,499,887	3,013,478
Total Scaffold Length (bp)	2,522,725,272	2,517,431,033	2,344,125,543	2,343,495,360	2,349,643,860	2,365,556,663
Captured Gaps	610,367	230,359	165,200	35,158	38,149	37,385
Max Gap (bp)	191,507	190,436	20,486	20,068	98,050	97,673
Mean Gap (bp)	798	873	355	954	1,040	934
Gap N50 (bp)	2,693	5,899	2,422	4,031	4,698	5,036
Total Gap Length (bp)	487,315,337	201,129,911	58,678,630	33,532,268	39,680,768	34,917,415

On the other hand, for improving the public Chinese hamster genome assembly, we had to use a slightly different approach due to the low fraction of gaps (2.5%) in the public genome. In comparison, the N50 scaffold length was lower (1.6 Mbp) compared to 2.6 Mbp for the public genome. We used a re-scaffolding strategy to improve the public Chinese hamster assembly. However, we first gapclosed the assembly prior to re-scaffolding. The fraction of gaps were reduced from 2.5% to 1.4% after gapclosing with the in-house short insert and long insert reads (Table 0-9). This gapclosed draft genome sequence was then re-scaffolded using the in-house long insert reads. AbySS standalone scaffolder was used with abyss-bwa as the mapping algorithm instead of abyss-map default mapping software (Jackman et al, 2013)

Ideally, the long insert reads from the public data could have been used to re-scaffold the in-house assembly to improve the scaffold lengths. However, this data was not deposited by the publishing group. By comparing the contiguity and gap length of the two draft assemblies (Table 0-9), the enhanced public genome was chosen to be used as the draft Chinese hamster genome.

7.4.3.4 Quality assessments of the Chinese hamster draft genome

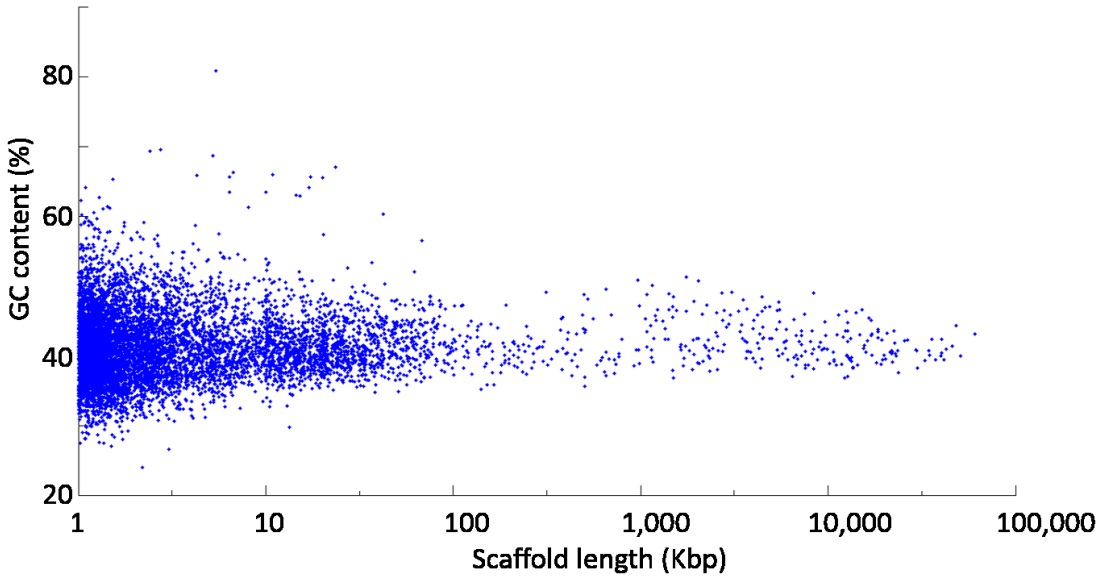


Figure 0-43: Scaffold GC content distribution shows no evidence of GC-bias in sequencing or assembly. The GC content of the scaffolds are evenly distributed around the average GC level.

Quality assessments were done on the draft Chinese hamster genome by aligning the EST contigs using a spliced alignment software BLAT (Kent, 2002). Out of the 353,445 EST contigs, almost 90% of the contigs, 315,432 contigs, aligned to the genome. The transcriptome assembly also contained the contigs from the transgenes in the CHO cells. It is obvious that the transgenes will not align to the Chinese hamster genome.

Most sequences obtained from modern sequencing technologies, including Illumina, may have a bias towards higher GC content. Such bias can be identified from analyzing the GC content of longer scaffolds. The median GC content of the scaffolds was 42%. The GC content profile is relatively flat at ~42% across all scaffold lengths, indicating unbiased sequencing and assembly and absence of GC-bias in sequencing (Figure 0-43).

Another way to validate the genomic scaffolds is by analyzing the gene synteny of Chinese hamster to mouse. The CHO EST contigs that had mouse annotation were aligned to the Chinese hamster genome using BLAT. An EST may have multiple hits (alignments) to the Chinese hamster genome scaffolds with varying degrees of confidence. For

simplicity, only the most confident hit for each EST was listed and termed as the "best unique hit". We first calculated a "score" for each hit, based on the difference between the total matches and mismatches within the hit (matches-mismatches). For each EST, the hit with the highest score was selected as the best unique hit. Each EST only had one such hit, hence they were "unique". Hits lower than 500 bp in length were discarded for this analysis. The position of the gene on mouse chromosome was plotted against the position in Chinese hamster scaffolds. A couple of examples are shown where synteny can be observed between Chinese hamster and mouse (Figure 0-44).

Most genes on Chinese hamster scaffold map to mouse genome contiguously. The close synteny similarity is suggestive of high quality scaffolding (Figure 0-44A). In the example on the lower panel, genes on Chinese hamster scaffold map to many mouse genome loci, specifically four different mouse chromosomes (Figure 0-44B). This is not necessarily indication of wrong scaffolding, because there could be many rearrangements between Chinese hamster and mouse, considering the evolutionary distance between them.

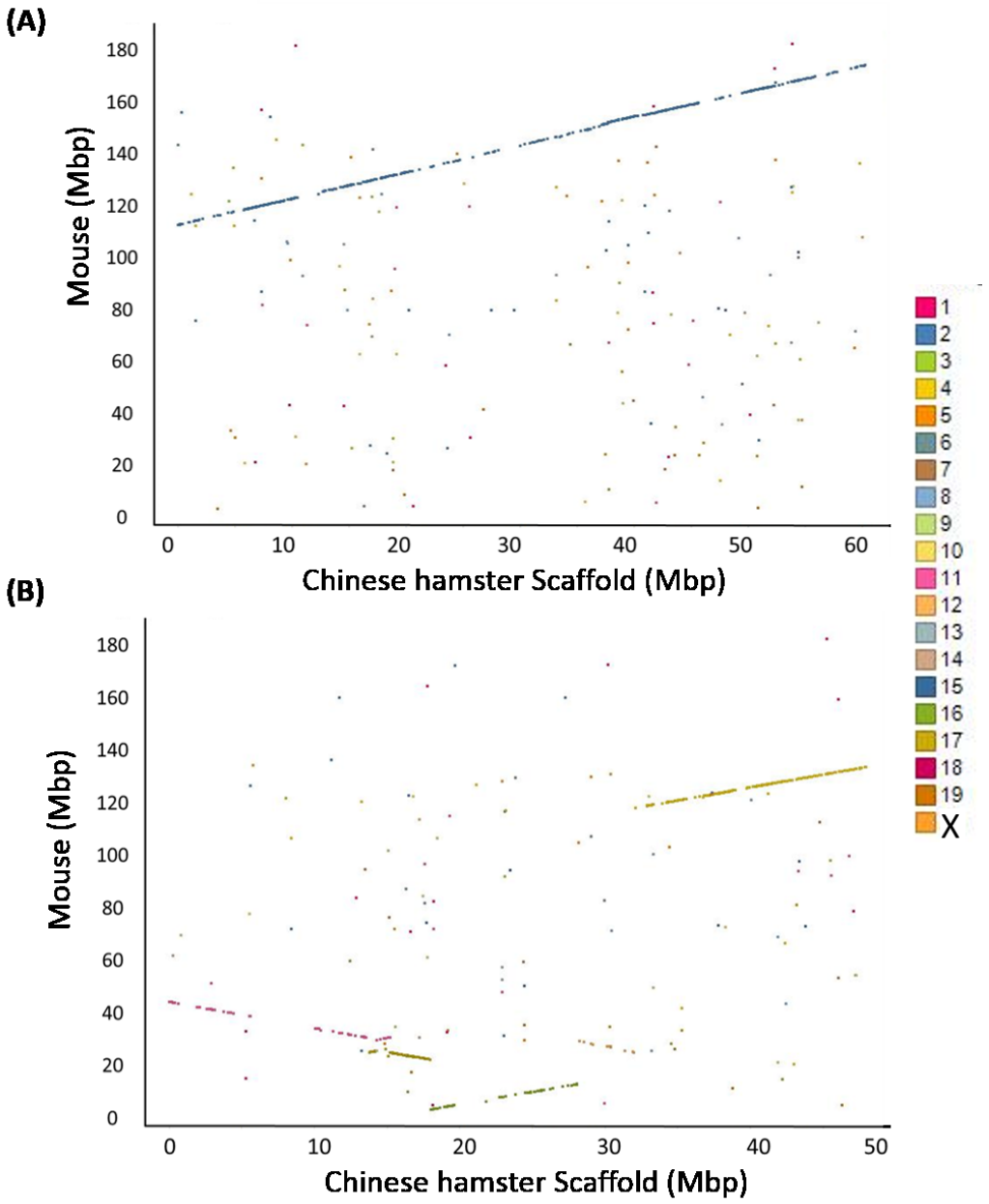


Figure 0-44: Quality assessment using synteny between mouse and Chinese hamster. The order of genes on mouse genome was compared to the order of genes in the assembled Chinese hamster genome. Conserved regions, shown in (A) exhibit high similarity in gene order between Chinese hamster and mouse. In this case, the gene ordering in a 60 Mbp region of mouse chromosome 2 is retained almost entirely in a Chinese hamster scaffold. In a region with lower conservation, such as shown in (B), a Chinese hamster scaffold shows hits to four different chromosomes in mouse. Within each hit, the genes are in the same order as mouse.

All the scaffolds that had more than 40 genes were assessed for the presence of such synteny. Out of 147 scaffolds tested, a total of 122 scaffolds showed good synteny conservation, similar to Figure 0-44A.

7.4.3.5 Features of the Chinese hamster genome

The average GC content in Chinese hamster genome is 41.4%, very close to the average of 42% in mouse. The GC content distribution of the genomes were also compared. The genome was divided into non overlapping windows of 20 Kbp and the GC content within these windows was calculated. The GC content distributions of Syrian hamster, Chinese hamster, mouse and rat genomes are very similar. GC content distribution in human is quite different from the rodents, with more windows showing lower GC content (Figure 0-45).

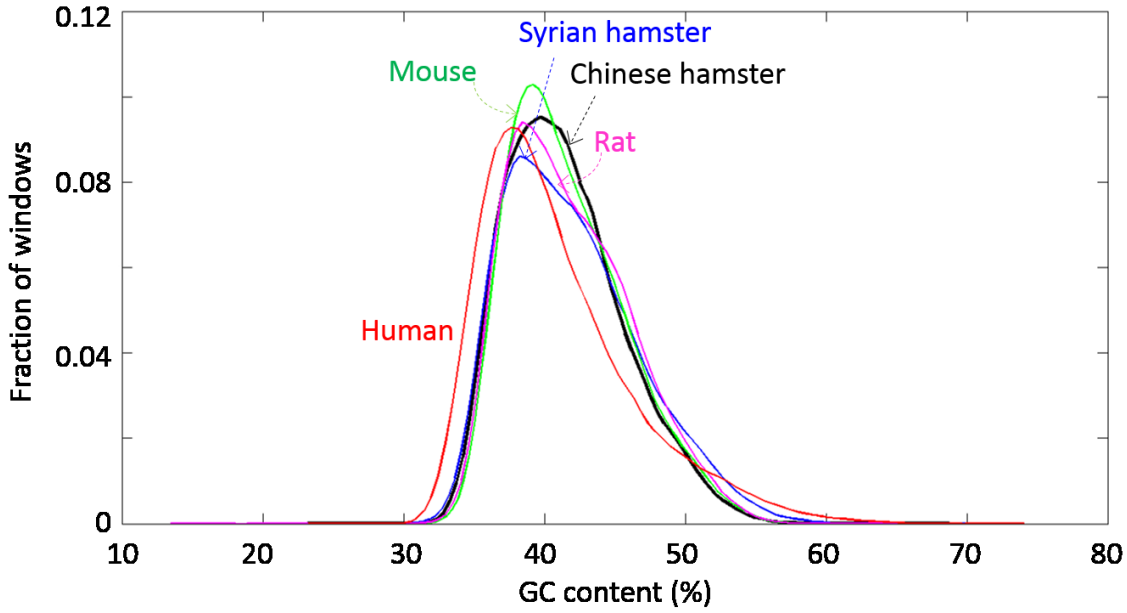


Figure 0-45: GC content distribution for human, mouse, rat, Syrian hamster and Chinese hamster. The genome was divided into non-overlapping windows of 20 Kbp each and the GC content of each window was calculated. The distribution of this GC content is plotted on the graph. The GC content of Chinese hamster is very similar to mouse and Syrian hamster, and then to rat. Human genome has overall lower GC content compared to the rodent species.

The EST contigs having Ensembl mouse annotation representing 24,516 unique Ensembl mouse genes were aligned to the Chinese hamster genome. A total of 920 scaffolds out of 7,873 scaffolds contained all the transcriptome contigs. The remaining scaffolds either did not possess any gene or possessed only fractions of the other genes. The distribution of the genes among the scaffolds reveals that many of the scaffolds are very dense in genes, whereas others are rather sparse. About 130 scaffolds had more than 50 genes per scaffold (Figure 0-46).

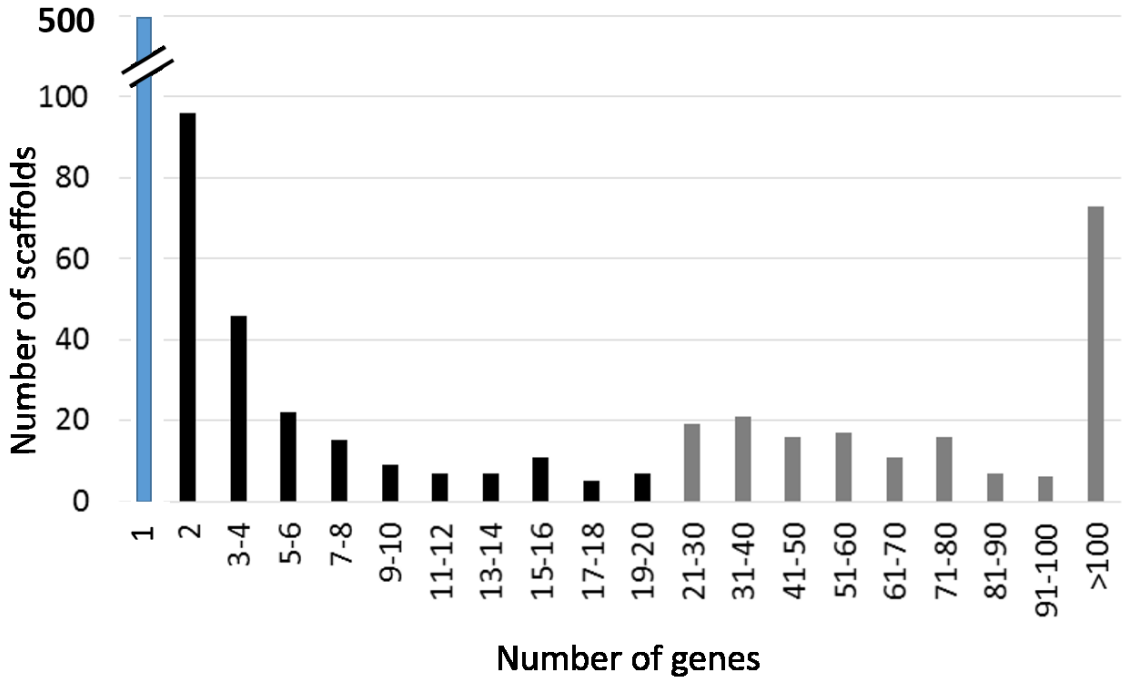


Figure 0-46: Distribution of genes over the Chinese hamster genomic scaffolds (bins are not uniform). Most genes are present on few scaffolds. About 500 scaffolds have one gene per scaffold.

Using the RepeatMasker program, the genome was also scanned for repetitive DNA sequences (Tarailo-Graovac & Chen, 2009; Tempel, 2012). Repeat sequences in the genome were searched against the repeat sequences in the genomes of the rodentia class in Repbase. A total of 35.4% of the genome was found to have evidence of repeats. This is comparable to the repeat content in mouse which is 38%, and is much lower than that in human (46%). Long interspersed nuclear elements (LINEs) were overrepresented in the

repeat sequences, comprising almost 14% of the genome. Short interspersed nuclear elements (SINEs) and long terminal repeats (LTRs) were the next most abundant, each represented 9% of the genome. While SINEs and LTRs are present at similar rates in mouse and Chinese hamster, the content of LINEs is estimated to be about 19% of the mouse genome as compared to 14% identified in the Chinese hamster genome.

Table 0-10: Repeat sequence analysis in Chinese hamster genome. Repeatmasker was used to identify repetitive sequence elements in the genome by comparing the sequences to known repeats in rodent genomes. Similar to mouse, LINE sequences are a major component of the assembled repears, followed by SINEs and other LTR elements.

		Number of elements*	Length occupied (Mbp)	Percentage of sequence
SINEs		1,648,193	216	9.2 %
	Alu/B1	655,144	77	3.29 %
	B2-B4	712,243	113	4.79 %
	IDs	94,050	7	0.29 %
	MIRs	116,086	14	0.6 %
LINEs		607,822	323	13.77 %
	LINE1	538,196	311	13.24 %
	LINE2	57,410	11	0.45 %
	L3/CR1	10,530	2	0.07 %
LTR elements		659,358	202	8.58 %
	ERVL	89,244	26	1.09 %
	ERVL-MaLRs	342,673	103	4.39 %
	ERV_classI	47,873	14	0.61 %
	ERV_classII	175,813	58	2.46 %
DNA elements		149,677	31	1.3 %
	hAT-Charlie	97,977	189	0.8 %
	TcMar-Tigger	29,812	7	0.3 %
Unclassified		27,253	129	0.55 %
Total interspersed repeats			785	33.4 %

* most repeats fragmented by insertions or deletions have been counted as one element

However, the number of LINE elements are similar between Chinese hamster and mouse. This could be potentially caused by the difficulty in assembling longer repeats,

leading to incomplete assembly or large number of gaps between the genomic scaffolds containing the LINE elements (Table 0-10).

7.4.4 Conclusions

The decreasing costs of high throughput sequencing has brought in a major breakthrough for genomic studies. Progressively more genomes are getting sequenced, and large amounts of data are being generated by the day. Considering the importance of CHO cells in bioprocessing, it was very critical to generate and use genomic information for improving their performance. Several parallel efforts, towards sequencing the Chinese hamster genome between 2011 and 2013 made a lot of genomic sequence available. Our group also sequenced the Chinese hamster genome. We undertook the efforts to integrate the information from all these sources to generate a comprehensive and high quality reference genome for Chinese hamster. Although a large number of good softwares for assembling genomes are available and being improved on a regular basis, a major challenge in this process was the unavailability of software for conducting merging of such genome assemblies. Our rationale of approaching this challenge will certainly aid other researchers to guide similar efforts in the future.

The availability of genome sequence for Chinese hamster opens up a plentitude of avenues for research. CHO cells have recently been referred to as a ‘quasispecies’ because of the extent of genomic variability observed among them. The availability of the genome can enable the detailed study of the genomic variability among different CHO cell lines. It will also help in estimating the genomic variability in culture. Regulatory bodies require therapeutics producing companies to use stable and clonal cell lines for production to avoid variability in product quality. Despite tightly controlling the process, the variability in product titer and quality is still high. Studying the genomic and transcriptomic changes in these cells at different stages of production can give important cues to intervene the processes and make them more consistent.

Genome editing tools like ZFN, TALEN or CRISPR/Cas9 systems can be used to knockout genes that affect protein quality or efficacy. For example, the knockout of the

fut8 gene leads to defucosylated antibodies in CHO cells, and leads to improved efficacy of antibodies through increased ADCC activity. Genome editing tools can potentially be used to target transgenes to genomic regions that are stable and transcriptionally active. These regions can be discerned from transcriptomic studies.

The applications of genomics in CHO cell bioprocessing are immense. The present state of efforts have just scratched the surface of the discovery space. The Chinese hamster genome sequence holds great promise for heightening the state-of-art practices in recombinant protein therapeutics production in CHO cells.

7.5 Integration site analysis- method development

7.5.1 Introduction

The duration of the process of developing a stable cell line for the production of a protein therapeutic drug candidate is 4-12 months. After every step in this process, the cells have to be screened in large numbers to ensure the stability of the recombinant gene expression. In most cases, tens of copies of the transgene are randomly integrated into the host cell's genome. After the amplification step using MTX or MSX treatment, many of these copies get amplified to hundreds, or even thousands. After the selection pressure is withdrawn, the cells lose some of these copies. After the cells stabilize the gene copy numbers, they are selected for further processing. Often in this process, the cell lines become unstable because of the silencing of the transgenes or because of the loss of copy number. Instead of relying on random integration, by targeting the transgene to an active locus, one can ensure stable and high expression of the transgene. Moreover, a single copy of the transgene may be sufficient to achieve high expression, thus avoiding the amplification process entirely. The elimination of the amplification step will permit a major compression of the time required for cell line development from a few months to a few weeks. For example, the transgene of a single copy high producing cell line can be swapped with transgene of another drug candidate by a simple transfection, which may be followed by a sub-cloning step. One such example is already reported, where enhanced expression and stability regions have been identified by screening a large population of sites (Chen et al, 2010).

The employment of amplification-free methods in bioprocess necessitates widespread efforts towards cataloguing the stability and expression of various integration sites. Another approach is to identify the integration of site of some very stable, high expression, low copy cell lines in use in the industry. The ability to quickly analyze the integration site of cell lines will facilitate such analyses and enable the discovery of such favorable genomic sites. Such attempts are especially feasible today because of the availability of the Chinese hamster genome.

Current methods for integration site analysis require the knowledge of the vector region flanking the insertion site. This enables region-specific primer design, leading to the enrichment of the region containing the integration site. Such methods are possible for retroviral or lentiviral based gene transfection because the integration site for these systems is precisely known (Bryda et al, 2006; Schmidt et al, 2007). In plasmid transfections, even if the vector is linearized, the integration site can vary considerably from the cut-site. Hence, a new method capable of identifying these sites is needed. This study attempts to find and evaluate a few methods for identifying such integration sites.

In typical cell line development, an expression vector containing the transgene is linearized by restriction enzymes, and then transfected into CHO cells. The vector gets randomly integrated into the CHO genome, most probably at the cut-site. It is quite possible that the ends of the vector get chewed up prior to integration. Sometimes, non-linearized vectors are used for transfection, in which case, the site on the vector at which the integration occurs is not known.

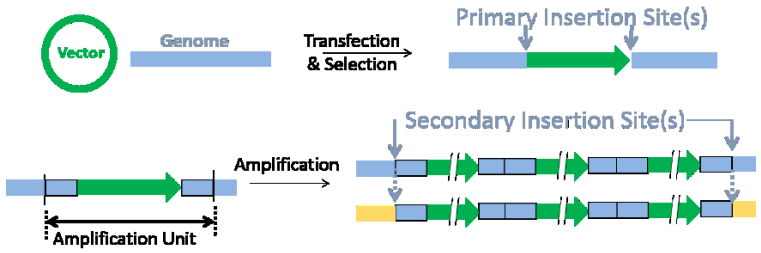


Figure 0-47: Primary and secondary insertion sites- When the plasmid vector is transfected into host cell, it randomly integrates at a certain site. This site is referred to as the primary insertion site. A cell may have many primary insertion sites. During the process of amplification, a part of region adjoining the integrated plasmid vector is co-amplified with the plasmid. This process may create many structural rearrangements. The amplification unit can also be transferred to different locations in the genome. This rearrangement creates secondary insertion sites. Each primary insertion site may have many corresponding secondary insertion sites.

A producing cell subjected to the standard cell line development process, possesses two types of integration sites. The site where the exogenous DNA integrates is called the

primary insertion site. A cell may have more than one primary insertion site. To make higher amounts of protein, the cells are subjected to DHFR-based transgene amplification in CHO cells that are deficient in DHFR function. By adding methotrexate, a DHFR antagonist, amplification of the region around the transgene integration site occurs. This region contains the plasmid vector and the surrounding genomic region and the primary integration is co-amplified. This co-amplified region may be duplicated and translocated to different genomic regions. These translocated sites are secondary insertion sites. Each primary insertion site may have many secondary insertion sites (Figure 0-47). The method developed in this study determines only the primary insertion sites.

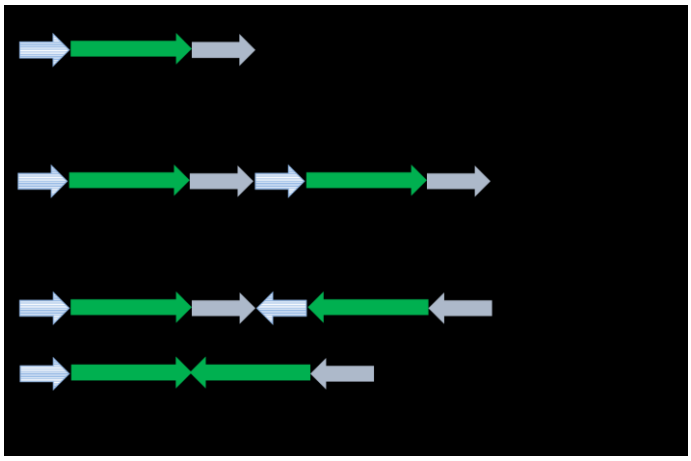


Figure 0-48: Possible scenarios of amplification. During the amplification step, many structural rearrangements occur. A simple amplification will involve direct amplification of the amplification unit in a head to tail arrangement. This amplification unit may be inserted in an inverted orientation compared to the primary insertion. During amplification, portions of the vector or genome may be chewed up by nucleases prior to integration, causing the deletion of some intervening sequence.

The actual amplification is very complex, and there may be many different ways that a genomic region may get amplified. A few possible scenarios are presented here to illustrate this complexity. The simplest amplification is the head to tail arrangement of amplification unit in tandem. Other possibilities include a head to head arrangement of the amplification unit, similar to an inverted repeat; or a head to head arrangement of

amplification unit along with the deletion of some of the intervening sequence (Figure 0-48).

7.5.2 Results and discussion

We used a cell line with a known integration site to study different insertion site analysis methods. The integration site of these cell lines were determined *a priori* using whole genome sequencing. A high producing cell line with high vector copy number (300), and a high producing cell line with low vector copy number (15) were used for the study. We tested a few popular methods commonly used to detect integration sites in lentiviral transfectants. These methods did not work in our system (Bryda & Bauer, 2010; Zou et al, 2003). Two of the methods that were successful for insertion site analysis were: (1) sequencing adaptor ligation-based PCR; and (2) physical separation using sequence capture, that is, solution-based methods for pull-down with biotinylated probes using streptavidin-coated beads.

7.5.2.1 Sequencing adapter ligation based PCR for targeted sequence enrichment

The genomic DNA was sheared and selected for a specific size, and then the DNA fragments were ligated with standard Illumina sequencing adapter. One primer from the vector sequence and the other primer from the adapter sequence, were used to amplify the conjunction sequences using PCR.

Controlled shearing of the genomic DNA was achieved by sonication. The shorter fragments were discarded. Three types of DNA fragments were obtained after shearing, (1) fragments containing the vector only, (2) fragments containing the genome only and (3) fragments containing the junction between the genome and the vector. Our objective was to enrich the fragment of the type (3) (Figure 0-49).

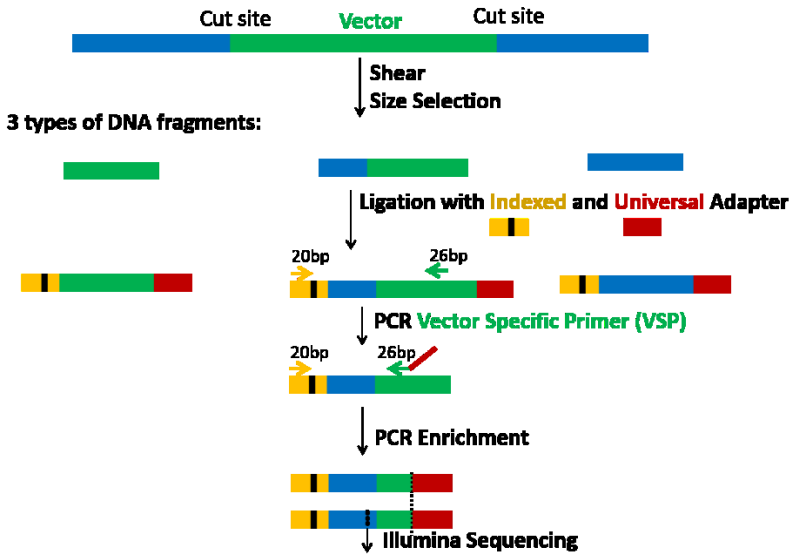


Figure 0-49: Schematic of the sequencing adaptor ligation based PCR for sequence enrichment. The sheared genomic DNA from the subject cell line will have three kinds of fragments – (1) containing only genomic DNA, (2) containing only vector DNA or (3) containing the vector genome-junction sequence. The objective is to enrich this pool of fragments in fragments of type (3) for integration site analysis. The Illumina sequencing library is first prepared by ligation of sequencing adapters to either ends of all fragments. Then, a PCR primer specific to adapter, coupled with a vector-specific PCR primer are used to conduct PCR. The PCR product will thus, be enriched for vector-containing regions, specifically regions of types (2) and (3). Another PCR with a nested primer is done along with the adapter-specific primer to enrich the pool further for vector-containing sequences. These fragments are then sequenced in an Illumina Mi-Seq machine.

In the next step, all the DNA fragments were ligated with the standard Illumina indexed adapter and universal adapter. To enrich the specific DNA fragments containing the vector genomic DNA junction, a 20 bp primer hybridizing to the indexed adapter, and a primer containing 26 bp from the vector and also linked to the universal adapter were designed. The universal adapter is necessary for the subsequent sequencing steps. The resulting PCR reaction is size-selected to remove smaller fragments, as they may not be long enough to sequence into the genome to identify the integration site confidently, and hence, will cause a loss of sequencing depth in the later steps. Since the PCR will be successful only if the primer binds to the vector region; by sequencing long enough, one can identify the integration site by Illumina sequencing.

One important design criterion for using this method is the primer design, because, the detection of the integration site is limited to the location of primer. Plasmids are usually linearized prior to transfection, and hence there is a high probability of the integration site to occur at or near the cut site. So, our primer design focused more on the regions close to the cut-site.

7.5.2.1.1 Sequencing adapter ligation based PCR method tested in high transgene copy number cell line

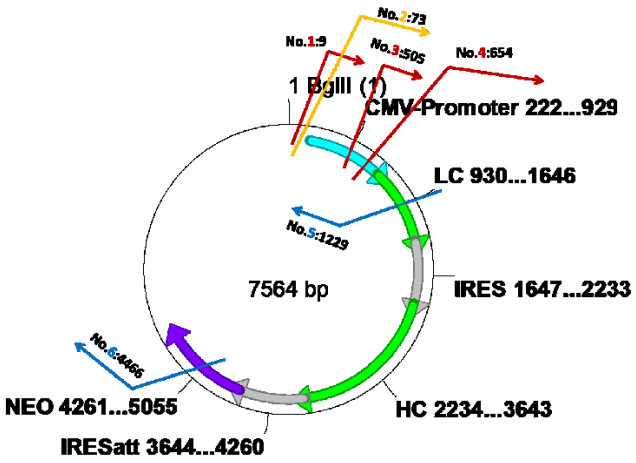


Figure 0-50: Locations of the 6 known integration sites in the vector used for the high copy cell line are indicated on the plasmid DNA. Most of the integration sites are close to the cut site (position 1). Some of the integration sites (5 and 6) form truncated transcripts of the transgenes, and are thus, non-functional for protein production.

A cell line with a known integration site was chosen to develop the methodology. The 6 known integration sites are mapped to the plasmid shown in the figure (Figure 0-50). The position 1 on the vector is the cut-site. Most integration sites (4 out of 6) occur within the first 900 bp of the plasmid cut-site, as we had expected. The direction of the arrow represents the portion of the plasmid that was retained at the integration site. Integration sites 5 and 6 are non-functional because a major portion of the transgene sections were truncated. The primer design was biased to preferentially span the region upstream of the light chain gene.

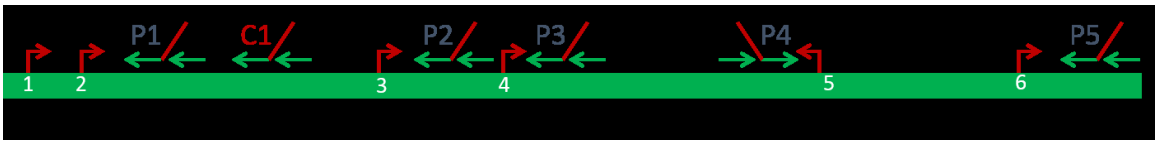


Figure 0-51: Primer design for the enrichment of the six known integration sites. The positions in the vectors are the number shown below the green bar (not to scale). Numbers on the green bar are the respective integration sites. Primers P1 through P5 are used to probe for the known integration sites. Primer C1 is a control primer as there is no integration site in proximity to the portion of the plasmid that can be amplified by the primer.

The first primer P1 was designed close to the cut site (< 100bp within the cut site). We designed three more primers C1, P2 and P3 in the region between the cut-site and the CMV promoter. This is because if any integration site exists beyond the CMV promoter, it will render the transgene inactive. So, we added more primers upstream of this region. We also designed two more primers P4 and P5 in order to enrich the other two integration sites. The primers were chosen such that there is one primer for every 200 bp. This made sure that any integration site within this region will show up in at least one PCR reaction. C1 is called a control primer because the results from that primer do not show any integration site. This was expected because there was no known integration site in close proximity to primer C1 (Figure 0-51). The PCR reactions from all the six primers were loaded into two MiSeq lanes. The read length obtained was 250 bp. Approximately, 12 million single-end reads were obtained from each lane.

Following sequencing, bioinformatics processing was done to identify the integration site. Briefly, the reads mapping partially to the vector as well as the genome were filtered out, and the number of reads supporting the same vector-genome junction are calculated. The Smith Waterman based algorithm bwa-sw tool was used for mapping, in order to allow for partial hits to the genome and vector (Li & Durbin, 2010). We counted the frequency of the border base of vector-genome junction for all the integration sites. The site with the highest frequency is the most enriched fragment and is a real integration site. The number of reads supporting the junction, will be hereforth referred to as the "*depth*" of the integration site. The first bioinformatics step is the trimming of sequences from the

Illumina adapters from the reads. After this step there was almost no loss in the number of reads, which is indicative of good library preparation. The Chinese hamster genome is about 2.5 Gbp in length. The integration site represents only a very small fraction of this genome. Even though we have a sizeable enrichment in the integration site region after PCR, it is unavoidable to have some genomic background. Out of 21.4 million reads, 4.7 million reads mapped completely to the genome because of background carry over. Out of the remaining 16.7 million reads, 12.2 million reads mapped entirely to the vector, i.e. having no genomic sequence. Such reads cannot be avoided, more so, in cell lines that have high copy number of the plasmid vector. The remaining 4.5 million reads were mapped to the vector and the genome to identify vector genome junctions. Out of these, 82 thousand reads were utilized towards identifying the integration sites. The remaining reads were not long enough to confidently contribute to the count of integration site and were then discarded.

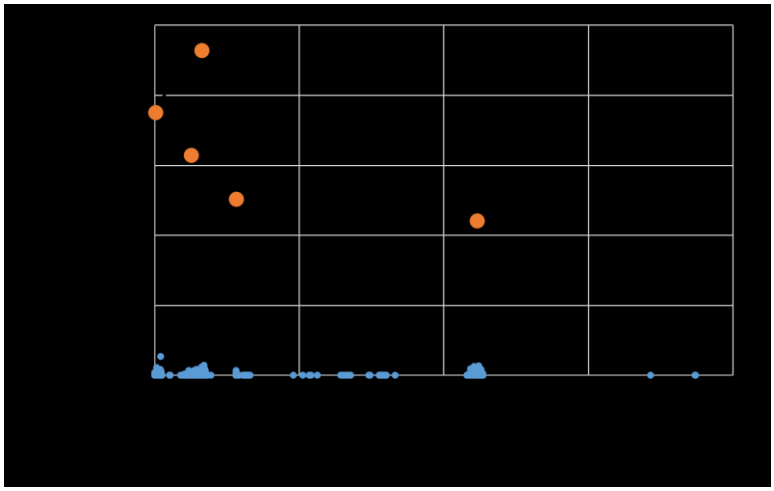


Figure 0-52: Depth of integration site from bioinformatics analysis for the sequencing adaptor ligation based PCR method. The data points marked 1, 3, 4, 5 and 6 are the data points corresponding to the integration site. Most integration sites are identified to very high depths, separating it from the background data points.

For each of the identified vector-genome junctions, the depth was calculated and plotted across the length of the vector. The integration site nos. 1, 3, 4, 5 and 6 were identified at much higher depths compared to the depth of the other junctions that formed the background. Integration site 2 was not identified. Even in whole genome sequencing,

the depth of the integration site 2 was very shallow, probably because it came from a small sub-population of cells (Figure 0-52). Since the cells were confirmed to be clonal, this observation is puzzling. It is possible that the entire population possessed this site initially, but slowly lost this site during sub-culture, and only a small population of cells with this integration site remain.

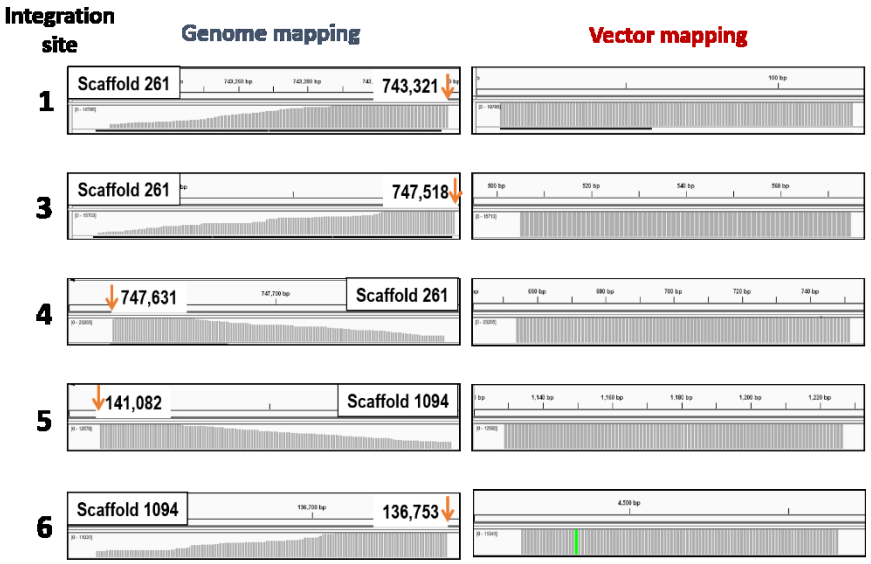


Figure 0-53: The pileup of reads mapping to genome and vector sections of the five integration sites identified by the sequencing adaptor ligation based PCR method. Genome mapping is shown on the left and vector mapping is shown on the right. The vector mapping shows a sharp boundary at the integration site, as well as the where the primer was designed. The genome mapping shows one sharp boundary at the integration site and a trailing boundary on another end. This is another confirmation of the presence of the integration site.

It is necessary to confirm if the integration site is not a PCR artefact. This is done by inspecting the insertion site for the absence of such artefacts. A real integration site would show blunt mapping to genome on one side of all the reads corresponding to the vector genome junction, and a trailing of depth on the other side in a staircase-like pattern. The mapping to the vector will be blunt on both the sides of the integration site. Mapping to the vector will also be blunt on the other side where the PCR primer is bound. For all the integration sites identified in the chart, the nature of vector and genome mapping was as expected indicating the accurate identification of the integration sites (Figure 0-53).

7.5.2.2 Physical separation- Solution phase sequence capture of insertion site

In solution-phase sequence capture, DNA fragments containing the target sequence are hybridized to biotin-labeled oligo probes (baits), which are subsequently pulled down by streptavidin-coated beads. The captured sequence can be sequenced by traditional Sanger or high-throughput sequencing methods (Figure 0-54).

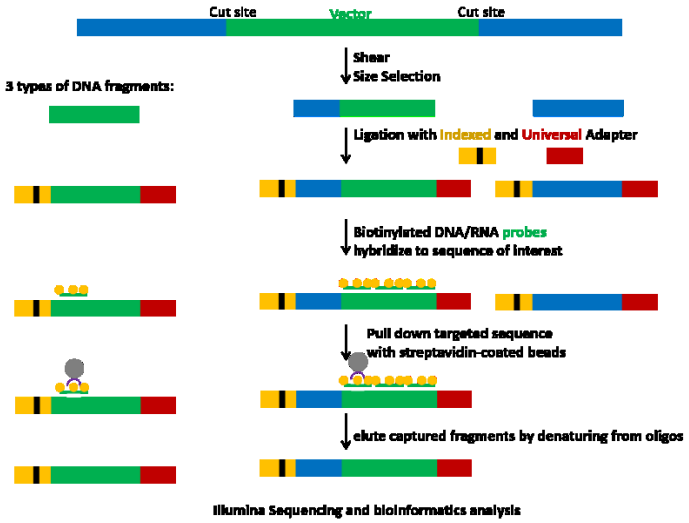


Figure 0-54: Schematic of the solution phase based sequence capture method for insertion site analysis. . The sheared genomic DNA from the subject cell line will have three kinds of fragments – (1) containing only genomic DNA, (2) containing only vector DNA or (3) containing the vector genome-junction sequence. The objective is to enrich this pool of fragments in fragments of type (3) for integration site analysis. The Illumina sequencing library is first prepared by ligation of sequencing adapters to either ends of all fragments. Biotinylated baits are constructed tiling the entire plasmid. The targeted sections of the genomic DNA are pulled down by the streptavidin-coated magnetic beads. These beads are separated from the solution using a magnet, and the captured sequences are eluted by denaturation.

Two cell lines were used to test the effectiveness of the sequence capture method to identify the integration sites. One has a high copy number of vector (~300) with about 6 known integration sites, and the other has a low copy number of vector (~15) with 3 known integration sites. The integration sites for these two cell lines were determined using whole genome sequencing.

7.5.2.2.1 Sequence capture for high transgene copy number cell line

Similar to the previous discussion on primer design for Adapter PCR-based method for integration site analysis, the bait design for sequence capture-based integration site analysis also emphasized the first 400 bp of the vector. This is because, there is a higher probability of the integration site occurring close to the cut-site. Since, it is quite possible that the insertion occurred within another part of the vector, we also designed a few baits tiling the entire vector. Baits were designed with a lower density accounting for the lower probability of its occurrence. Therefore, 10 times more baits were designed at the first 400 bp of the vector compared to the remaining vector length.

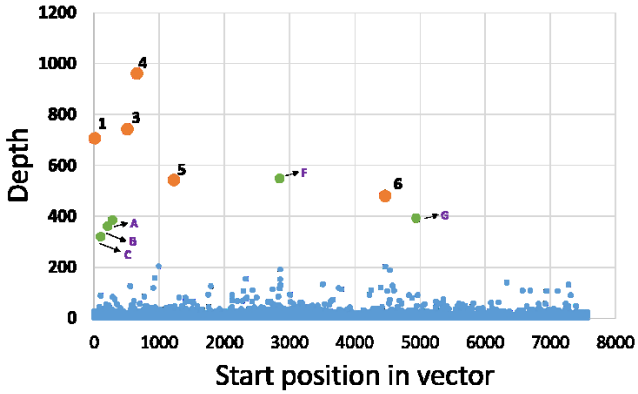


Figure 0-55: Depth of the integration sites detected by solution phase sequence capture technique. The data points marked 1, 3, 4, 5 and 6 are indicated on the graph. In this case, some other integration sites were also detected at higher depths above background. These are indicated as A, B, C, F and G on the figure.

Using the same bioinformatics methodology outlined in the previous section, several vector-genome junctions were identified. In this case, we could easily identify integration sites 1, 3, 4, 5 and 6 because of their stark difference in the detection depth from that of the background. Interestingly, even in this case, the integration site 2 was not detected. In addition to these known integration sites, other vector-genome junctions A, B, C, F and G were identified.

We further looked into the local mapping of the vector and genome regions at the candidate integration sites. For sequence capture, the mapping to genome should be blunt at the junction site for all the reads mapping to that region, and a trailing depth ladder on the other side. Similarly, mapping to the vector should also show blunt high depth at the

junction of the integration site, and a trailing depth like a staircase for the other side of the reads mapping to the vector. For each of the integration sites, we observed such a trend with blunt depth at the junction site and a trailing staircase-like depth on the other side for both the vector and the genome mapping.

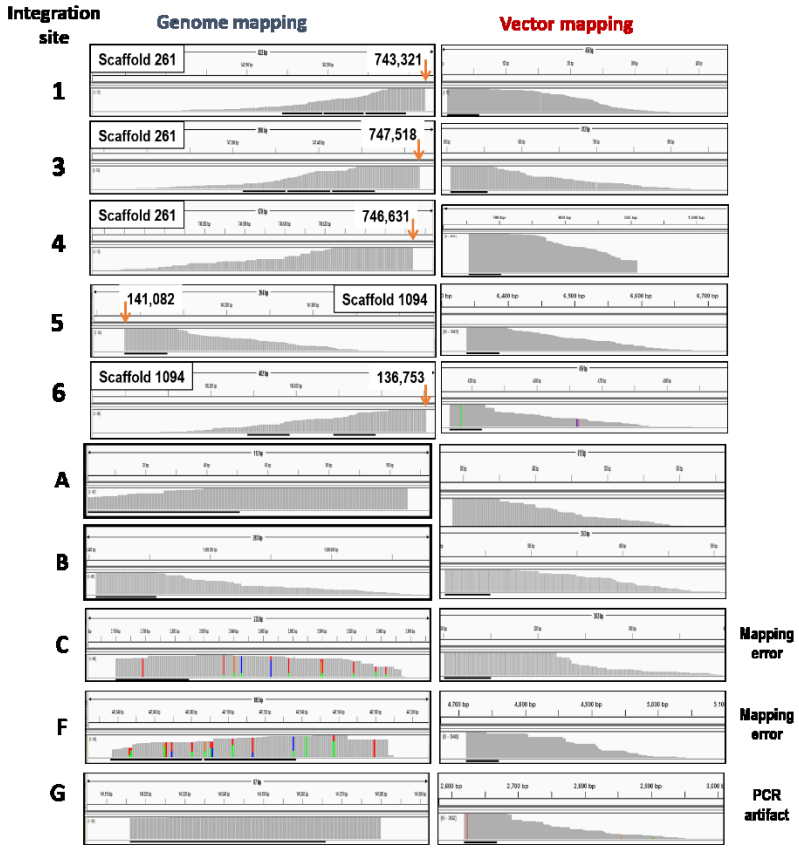


Figure 0-56: Pileup data of the captured sequences detected by the solution phase sequence capture method. Mapping to genome is shown on the left and mapping to vector is shown on the right. The integration sites 1, 3, 4, 5 and 6 show the expected mapping pattern with a sharp boundary at the integration site, and a trailing pattern on the other end for both vector and genome mapping. This confirms that the site detected has not been affected by PCR bias. Sites A, B and G may be other potential integration sites that were not detected earlier from whole genome sequencing. Sites C and F do not exhibit good quality mapping at the genome, thus dismissing the presence of these two integration sites.

Similarly, for the novel vector genome junction sites A, B, C, F and G, we looked at the mapping patterns to determine if they could be real integration sites. For site A and B, the mapping patterns seem to be real. They may be additional novel integration sites

that were not detected in the whole genome sequencing. For sites C and F, the mapping results showed a lot of mismatches in the genome mapping. So, it probably not a real integration site. For site G, the double blunt end of genome mapping indicates that it is a PCR artefact and may not be a real integration site.

7.5.2.2.2 Sequence capture for low transgene copy number cell line

The plasmid vector was not linearized prior to transfection in the low transgene copy number cell line. Hence, the biotinylated baits were designed equi-spaced along the length of the vector (every 120 bp) with 1X tiling. Bioinformatics analysis was performed as described previously.

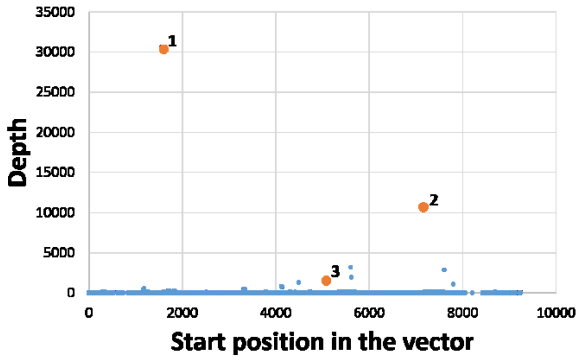


Figure 0-57: Depth of the insertion sites detected by solution based sequence capture method in the low copy cell line. Integration sites 1 and 2 are detected at high depths above background. However, integration site 3 is represented at lower depth relative to background. Even from whole genome sequencing, the depth of integration site 3 was very low compared to the other two integration sites.

The three known integration sites 1, 2, and 3 for the low-copy cell line were identified from the high contrast in depth compared to the background (Figure 0-57). For quality check, the mapping to vector and the genome were analyzed. As expected for a real integration site, the reads showed blunt mapping pattern at the integration site and a trailing pattern on the other side of the integration site. This was observed for mappings to the vector and the genome indicating that these integration sites are real (Figure 0-58).

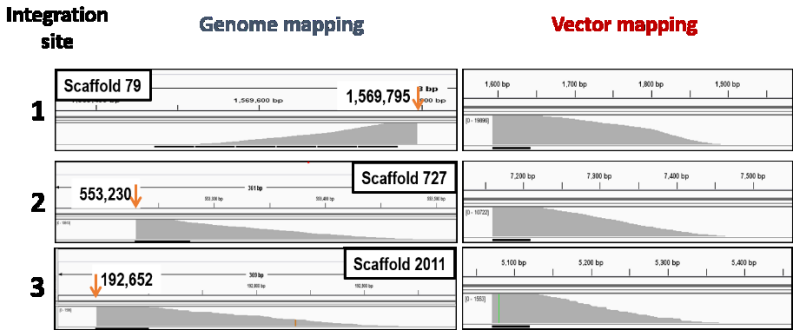


Figure 0-58: Pileup of reads at the vector and genome regions for the three integration sites detected by the solution phase sequence capture method for the low copy cell line. Genome mapping is shown on the left and mapping to vector is shown on the right. As expected, all three show sharp boundary at the integration site and a trailing pattern on the other end.

The neighboring regions near the integration site can play an important role in interacting with and regulating the expression level of the transgene. A summary of such immediate neighboring regions for all the integration sites is presented in Table 0-11. The absolute detection depths for the low copy cell line is higher than that for the high copy cell line. The lower vector amount in the low copy cell line has potentially reduced the background. Most of the baits captured the genomic region with the integration site, and only a few captured the vector background. Since both the cell lines are high producers (40 pg/cell/day), the integration sites of these cells are potential sites that can be used for targeted integration.

Most of the integration sites are intronic or intergenic. The expression level of the corresponding genes that were interrupted by the integration, was obtained from microarray data with a normalized average expression value of 500. The average expression level from RNA-Seq data was approximately 200. All the genes that were interrupted by the plasmid integration had average expression levels. Potentially, highly expressed genes are essential for the cell's growth and survival, and hence, the interruption of its expression may prove to be harmful for the cell. The integration at a moderately expressed gene may not be very disruptive and yet have active chromatin positively impacting transgene expression.

Table 0-11: Description of the gene structure at the integration site for both cell lines. Since both the cell lines are high producers, the rate of transcription of the adjacent regions may be of relevance. Most integration sites occur with an intron of a moderately expressed gene. Wherever relevant, the gene expression values of those genes in CHO cells from RNA-Seq and microarray are indicated on the table.

Insertion site	Position Vector	Depth	Scaffold Number	Scaffold position	Insertion site description	RNA-Seq RPKM	Microarray Intensity
High copy cell line							
1	9	707	261	743,321	Etv6(TF), Intron	148	492
3	505	742	261	747,518	Etv6(TF), Intron	148	492
4	654	828	261	747,631	Etv6(TF), Intron	148	492
5	6339	543	1094	141,082	Intergenic	-	-
6	4466	480	1094	136,753	Intergenic		
Low copy cell line							
1	1594	30,360	79	1,569,795	Rc3h1, intron	245	454
2	7161	10,722	727	553,230	Vsp13b, intron	1325	391
3	5072	1,553	2011	192,652	Intergenic	-	-

7.5.2.3 Comparison of integration site detection approaches

For the high copy number cell line, the contrast between the background and the signal was higher in the adapter PCR-based method compared to the sequence capture method. It must be noted that the sequencing depth for the adapter PCR-based method was almost 4-times than that of the sequence capture. Sequencing adapter based enrichment methods rely heavily on primer design, requiring a discrete array of primers to probe all the integration sites. This method is more suitable when the vector is linearized prior to transfection. When the junction sequence is unknown, many primers have to be designed along the vector increasing the cost of the analysis. One advantage of this method is that it has a higher yield of specific sequences with insertion site, because PCR enriches the targeted sequence considerably. The success of PCR-based methods rely heavily on the primer design. In contrast, the sequence capture based enrichment method uses probes that

tile the entire vector, and a single experiment can probe for all possible integration sites. This method can be used even when vector is not linearized. A sequence capture library has to be developed for every vector used, which increases the capital cost of this method compared to the PCR-based method. Most of the IgG sequences are similar in sequence with a few changes. If the same backbone is used for most of the cell lines containing IgG sequences, a generic library can be constructed for wide applicability. It can also be made cost effective by probing the integration site for many cell lines in a single experiment.

7.5.3 Conclusions

Two methods based on PCR and sequence capture were developed and tested for integration site analysis. These methods were used to study the integration site of two high producing cell lines, one with high transgene copy, and the other with low transgene copy. Although the PCR-based method shows a high detection depth for the integration sites, the detection of the insertion site depends on its proximity to the primers. An array of primers need to be designed along the vector to probe for all the possible integration sites. It is more suitable to be used for cell lines created by linearized vector transfection, where the most probable region on the vector for the integration site is known and primer design can be restricted to that region. In contrast, the sequence capture based method exhibits lower detection depths, however, a single experiment can theoretically probe all the integration sites in the cell line.

Most of the detected integration sites were located in the intronic regions, while a few of them were located in intergenic regions. The expression levels of the genes intercepted by the transgene integration is close to the average expression level. Since both the cell lines used were high producers, the location of these integration sites are of special interest for future cell engineering efforts.

The integration site characterization methods will find applications not only for bioprocessing, but also in other studies where the transgene is randomly integrated in large genomic backgrounds.

Chapter 8: Conclusions and Future Directions

Exciting times lay ahead for cell culture engineering, being in the peak of the 'genomics' era. The past few years have seen the birth of a plentitude of genomic resources for Chinese hamster and CHO cells, which have also brought in a lot of new discoveries. A major part of this dissertation is dedicated to developing and improving genomic resources for CHO cells. Even though the genome and transcriptome sequences were available for use, the annotation of these sequence was lagging behind. This challenge was addressed by employing a rigorous homology-based annotation strategy expanding the repertoire of annotated sequences considerably. Such genomic information is extremely valuable for taking bioprocessing to the next stage of development.

In the past few years, bioprocessing has seen increasingly more focus on process intensification and decentralization. The number of drug candidates that need to be developed are increasing, and the only way to remain competitive is to be fast in delivery. For a long time, recombinant protein therapeutics have been an elite drugs, mostly available in the developed countries, and to the better-offs. With many of the older and effective drugs going off-patent, developing countries are beginning to invest heavily in producing these drugs in house, and marketing them at much lower costs. The biotechnology industry has begun to have a more worldwide presence, especially in the past few years. Therefore, especially now, it is essential to make processes more time-efficient, more streamlined and well defined to ensure reproducibility.

The easy access to genomic information can help address many of these very pertinent issues. The community is now motivated to understand the underlying scientific principles, to ensure reproducibility. In most cases, a platform process is adopted to ensure smooth and efficient process development and transfer. Despite this, deviations are often observed. A better understanding of the molecular changes can lead us to cues for avoiding deviations. For example, understanding the inherent variability in gene expression levels of key pathways relevant to the process can give us useful insights to control the process.

Variability in gene expression among CHO cells of varied origin was studied using an expression microarray. RNA-seq analysis was used to compare expression levels in

CHO cells with the physiological levels in Chinese hamster tissues. A few important genes responsible for product quality are found to be absent in a few cell types. De-silencing these genes or knocking-in such genes can ensure consistent product quality. This variance in gene expression among CHO cells creates the necessary genetic diversity required for isolating high producing CHO cells for the protein production.

These characteristics that make cells hyper-producers were studied by analyzing the transcriptome of the cells during the process of cell line development. A complete non-producer that has no machinery to secrete proteins, gets transformed into a super-producer with productivities rivalling professional secretors like plasma and liver cells. Following transfection and selection, cells are usually subjected to an amplification step to increase the transgene copy number. From the transcriptome analysis of cells undergoing this transition, we speculate that amplification potentially creates a burst of messenger RNA overburdening the cells' protein processing machinery possibly leading to the triggering of unfolded protein response. Only the cells that have sufficient machinery to process this large load of protein survive, leading to the selective isolation of high producers. The hyper-productivity traits largely involve growth signaling, mRNA and protein processing, cell cycle and energy metabolism. The leading genes conferring the hyper-productivity trait were combined with those identified from other studies, thus generating a hyper-productivity gene set. The size of the gene set is too sparse and requires more data for further refinement before it becomes practically usable for screening cells.

In addition to selecting high producing cells, it is also important to have a robust process to ensure high productivity in every bioreactor run. One of the outstanding issues is the production of lactate, a metabolite that is detrimental to cell growth and productivity. While some cultures have the propensity to consume lactate in the late stage and perform well, some others produce lactate even under identical processes. Lactate consumption was achieved by the reduction of glycolytic flux in the late stage of the culture. This was carried out by expressing a fructose transporter selectively in the late stage of the culture, while simultaneously switching the sugar source to fructose. The transporter was expressed using a novel dynamic promoter that drove expression in-sync with the cell's growth, enabling

selective uptake of fructose only in the late stage of cell growth. The productivity of this culture was improved compared to the control culture.

Consistency of performance also relies on the genetic stability of the clonal cell line selected for production. It is often observed that the integration site of the transgene plays an important role in determining the stability of the clone. Most practices still rely on random integration followed by intensive selection to identify a stable clone. In contrast, the state-of-art practices are moving towards a targeted integration approach where the transgene is directed to a predetermined stable and transcriptionally active genomic site. In order to identify such a site, a lot of integration sites are needed to be screened. Two methods for the identification of integration sites were optimized and evaluated. The integration sites of two stable and high producing cell lines were identified. The integration sites were found within the intron of a medium expression gene. More such sites need to be evaluated in the future, and a few candidate sites can then be used for targeted integration.

An ideal bioprocessing practice should involve a combination of all the efforts presented in this dissertation. A targeted transgene integration approach applied for genetically stable clones, followed by a more refined transcriptome-based cell screening to identify high producers will open the door for faster and controlled process development. The selection of a high producing clone needs to be coupled with a robust process to ensure maximum productivity.

8.1 Future directions

The next phase of genomics in bioprocessing should involve exploration of the epigenome in addition to the transcriptome and genome. The transformation of CHO cells from non-producers to extremely high producers entail colossal changes in the transcriptome profiles, which potentially arise due to epigenetic changes. These epigenetic changes, if identified, can be engineered to make higher producers. The genes showing massive changes in gene expression levels, or those that are highly variable, are probably

subjected to epigenetic modification. The epigenetic states of some of these genes that are relevant for bioprocessing should be explored.

One such gene is glutamine synthetase, which makes cells auxotrophic for glutamine. Glutamine is a metabolite that plays an important role in lactate metabolism. Since lactate metabolism is critical for bioprocessing, the epigenetic state of the glutamine synthetase gene is highly relevant. The upstream region of the gene has a dense and conserved CpG island. The frequency of obtaining glutamine independent cells after de-silencing by 5-azacytidine had drastically increased (Figure 0-59A). A major fraction of DNA at the upstream region of the glutamine synthetase gene was unmethylated in comparison to control liver DNA which was totally unmethylated (Figure 0-59B). Glutamine independent cells can be isolated and probed at the upstream CpG island for reversal of DNA methylation.

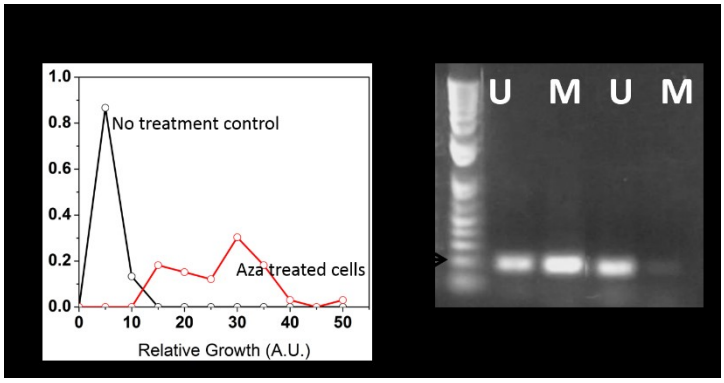


Figure 0-59: (A) Increased frequency of isolating glutamine independent cells on treating with 5-azacytidine, a de-silencing drug. (B) Methylation specific PCR showing a large fraction of methylated DNA at the upstream region of the glutamine synthetase promoter, where the PCR probes were designed.

Similarly, another gene α(2,6)-sialyltransferase shows no expression in the CHO cells tested in our study. This gene also has upstream CpG islands that can be probed for potential silencing. Sialic acid linkage to glycoform in humans is α(2,6) type, as compared to the α(2,3) type in CHO cells. This can be attributed to the highly expressed α(2,3)-sialyltransferase. Cells can be engineered to activate the endogeneous gene in order to obtain more human-like glycoform of the product.

Consistency in glycoform or product quality is one of the major roadblocks faced by the industry today. The extensive microarray data generated in this study can be used to study the bottlenecks of the glycosylation pathway. The historical mathematical models developed for the glycosylation pathway can be revamped by including more realistic enzyme levels estimates obtained from microarray data. Possible avenues for intervention can be explored.

The possibility of targeted genome integration is another exciting avenue that can be pursued in the future. Random single copy integration of GFP gene followed by sorting high GFP expressers can be used to identify hot spots. Further, the cells can be cultured for extensive periods of time. The cells that continue to express high GFP levels can be isolated by sorting. The integration site of these cells can be identified and catalogued. GFP can be exchanged for a therapeutic protein, and the productivity of the cell can be ascertained. By targeting to a transcriptionally active and genetically stable site, a single copy high producer can potentially be obtained. This cell line will be very valuable because the transgenes can be swapped in the future for reproducibly isolating high producing clones for other drug candidates.

Our hypothesis of the role of the amplification step in obtaining high producing cells can be further explored. The cells can be transiently transfected with a high concentration of mRNA of the product gene, and the frequency of isolating high producers can be evaluated. The expression profile of the genes can be probed and compared with that of the genes differentially expressed during the amplification step. This will give valuable hints for shortening the time-scale of cell line development.

On the other hand, dynamic cell engineering can also be exploited for improving cellular productivity. The experiment in mixed substrate culture using cells dynamically expressing glut5 transporter with switch back to glucose (Figure 0-34), can be reattempted with pH control in a more controlled environment like a bioreactor. Anti-apoptosis engineering will also benefit from the use of a dynamic promoter. In cells engineered for anti-apoptosis because of slow death, cultures last longer and perform very well and have higher titers. However, invariably high expression of these gene in the growth stages

interferes with cell cycle control, leading to slower growth rates. They often have to be engineered using inducible promoters so that their expression can be activated in the late stages of the culture.

We are just at the beginning of what can be achieved from application of genomics in developing productive, stable and reproducible processes. The genomics era will bring in innovative solutions for robust and efficient bioprocessing of the future.

References

(2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455:** 1061-1068

(2013) The Broad Institute. *Mesocricetus auratus Genome sequencing*, http://www.ncbi.nlm.nih.gov/bioproject/77669.

Abu-Absi S, Xu S, Graham H, Dalal N, Boyer M, Dave K (2014) Cell culture process operations for recombinant protein production. *Advances in biochemical engineering/biotechnology* **139:** 35-68

Abu-Absi SF, Yang L, Thompson P, Jiang C, Kandula S, Schilling B, Shukla AA (2010) Defining process design space for monoclonal antibody cell culture. *Biotechnol Bioeng* **106:** 894-905

Aggarwal RS (2014) What's fueling the biotech engine-2012 to 2013. *Nature biotechnology* **32:** 32-39

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403-410

Baik JY, Lee KH (2014) Toward product attribute control: developments from genome sequencing. *Current opinion in biotechnology* **30C:** 40-44

Baik JY, Lee MS, An SR, Yoon SK, Joo EJ, Kim YH, Park HW, Lee GM (2006) Initial transcriptome and proteome analyses of low culture temperature-induced expression in CHO cells producing erythropoietin. *Biotechnol Bioeng* **93:** 361-371

Bandaranayake AD, Almo SC (2014) Recent advances in mammalian protein production. *FEBS letters* **588:** 253-260

Barron N, Kumar N, Sanchez N, Doolan P, Clarke C, Meleady P, O'Sullivan F, Clynes M (2011) Engineering CHO cell growth and recombinant protein productivity by overexpression of miR-7. *Journal of biotechnology* **151:** 204-211

Barsoum J (1990) Laboratory Methods Introduction of Stable High-Copy-Number DNA into Chinese Hamster Ovary Cells by Electroporation. *DNA and Cell Biology* **9:** 293-300

Becker J, Hackl M, Rupp O, Jakobi T, Schneider J, Szczepanowski R, Bekel T, Borth N, Goesmann A, Grillari J, Kaltschmidt C, Noll T, Puhler A, Tauch A, Brinkrolf K (2011)

Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. *Journal of biotechnology* **156:** 227-235

Bengea CT (2008) Metabolic engineering of Chinese hamster ovary cells for reduction of lactate formation via siRNA-mediated knockdown of GLUT1 glucose transporter. M.Sc. Thesis, Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2014) GenBank. *Nucleic acids research* **42:** D32-37

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57-74

Bertolotti M, Yim SH, Garcia-Manteiga JM, Masciarelli S, Kim Y-J, Kang M-H, Iuchi Y, Fujii J, Vene´ R, Rubartelli A, Rhee SG, Sitia R (2010) B- to Plasma-Cell Terminal Differentiation Entails Oxidative Stress and Profound Reshaping of the Antioxidant Responses. *Antioxidants & Redox Signaling* **13:** 1133-1144

Bibila T, Flickinger MC (1991) A Structured Model for Monoclonal Antibody Synthesis in Exponentially Growing and Stationary Phase Hybridoma Cells. *Biotechnology and Bioengineering* **32:** 210-226

Birzele F, Schaub J, Rust W, Clemens C, Baum P, Kaufmann H, Weith A, Schulz TW, Hildebrandt T (2010) Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. *Nucleic acids research* **38:** 3999-4010

Boch J, Scholze H, Schornack S, Landgraf A, Hahn S, Kay S, Lahaye T, Nickstadt A, Bonas U (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326:** 1509-1512

Bonham-Carter J, Shevitz J (2011) A Brief History of Perfusion Biomanufacturing. *BioProcess International* **9:** 24-30

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou WC, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, Howard J, Hunt M, Jackman SD, Jaffe DB, Jarvis ED, Jiang H, Kazakov S, Kersey PJ, Kitzman JO, Knight JR, Koren S, Lam TW, Lavenier D, Laviolette F, Li Y, Li Z, Liu B, Liu Y, Luo R, Maccallum I, Macmanes MD, Maillet N, Melnikov S, Naquin D, Ning Z, Otto TD, Paten B, Paulo OS, Phillippy AM, Pina-Martins F, Place M, Przybylski D, Qin X, Qu C, Ribeiro FJ, Richards S, Rokhsar DS, Ruby JG, Scalabrin S, Schatz MC, Schwartz

DC, Sergushichev A, Sharpe T, Shaw TI, Shendure J, Shi Y, Simpson JT, Song H, Tsarev F, Vezzi F, Vicedomini R, Vieira BM, Wang J, Worley KC, Yin S, Yiu SM, Yuan J, Zhang G, Zhang H, Zhou S, Korf IF (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* **2:** 10

Brinkrolf K, Rupp O, Laux H, Kollin F, Ernst W, Linke B, Kofler R, Romand S, Hesse F, Budach WE, Galosy S, Muller D, Noll T, Wienberg J, Jostock T, Leonard M, Grillari J, Tauch A, Goesmann A, Helk B, Mott JE, Puhler A, Borth N (2013) Chinese hamster genome sequenced from sorted chromosomes. *Nature biotechnology* **31:** 694-695

Brown AJ, Sweeney B, Mainwaring DO, James DC (2014) Synthetic promoters for CHO cell engineering. *Biotechnol Bioeng*

Bryda E, Pearson M, Agca Y, Bauer B (2006) Method for detection and identification of multiple chromosomal integration sites in transgenic animals created with lentivirus. *Biotechniques* **41:** 715-719

Bryda EC, Bauer BA (2010) A restriction enzyme-PCR-based technique to determine transgene insertion sites. *Methods Mol Biol* **597:** 287-299

Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* **11:** 94

Burleigh SC, van de Laar T, Stroop CJ, van Grunsven WM, O'Donoghue N, Rudd PM, Davey GP (2011) Synergizing metabolic flux analysis and nucleotide sugar metabolism to understand the control of glycosylation of recombinant protein in CHO cells. *BMC biotechnology* **11:** 95

Candido EP, Reeves R, Davie JR (1978) Sodium butyrate inhibits histone deacetylation in cultured cells. *Cell* **14:** 105-113

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic

D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y (2005) The transcriptional landscape of the mammalian genome. *Science* **309:** 1559-1563

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, Forrest ARR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics* **38:** 626-635

Castro-Melchor M, Le H, Hu WS (2011) Transcriptome Data Analysis for Cell Culture Processes. *Advances in biochemical engineering/biotechnology*

Charaniya S, Karypis G, Hu WS (2009) Mining transcriptome data for function-trait relationship of hyper productivity of recombinant antibody. *Biotechnol Bioeng* **102:** 1654-1669

Chee Furng Wong D, Tin Kam Wong K, Tang Goh L, Kiat Heng C, Gek Sim Yap M (2005) Impact of dynamic online fed-batch strategies on metabolism, productivity and N-glycosylation quality in CHO cell cultures. *Biotechnol Bioeng* **89:** 164-177

Chen G, Babb R, Fandl JP. (2010) Enhanced expression and stability regions. Regeneron Pharmaceuticals, Inc., Tarrytown, NY (US), United States, Vol. US 8,389,239 B2.

Chevreux B, Wetter T, Suhai S (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Proceedings of the German Conference on Bioinformatics (GCB)* **99:** 45-56

Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, Bogdanove AJ, Voytas DF (2010) Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186:** 757-761

Chusainow J, Yang YS, Yeo JH, Toh PC, Asvadi P, Wong NS, Yap MG (2009) A study of monoclonal antibody-producing CHO cell lines: what makes a stable high producer? *Biotechnology and Bioengineering* **102:** 1182-1196

Clarke C, Doolan P, Barron N, Meleady P, O'Sullivan F, Gammell P, Melville M, Leonard M, Clynes M (2011) Large scale microarray profiling and coexpression network analysis of CHO cells identifies transcriptional modules associated with growth and productivity. *Journal of biotechnology* **155:** 350-359

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* **339:** 819-823

Cost GJ, Freyvert Y, Vafiadis A, Santiago Y, Miller JC, Rebar E, Collingwood TN, Snowden A, Gregory PD (2010) BAK and BAX deletion using zinc-finger nucleases yields apoptosis-resistant CHO cells. *Biotechnol Bioeng* **105:** 330-340

Crawford DR, Schools GP, Davies KJA (1996a) Oxidant-Inducible adapt15 RNA Is Associated with Growth Arrest- and DNA Damage-Inducible gadd153 and gadd45. *Archives of Biochemistry and Biophysics* **39:** 137-144

Crawford DR, Schools GP, Salmon SL, Davies KJA (1996b) Hydrogen Peroxide Induces the Expression of adapt15, a Novel RNA Associated with Polysomes in Hamster HA-1 Cells. *Archives of Biochemistry and Biophysics* **325:** 256-264

Cruz HJ, Moreira JL, Carrondo MJ (1999) Metabolic shifts by nutrient manipulation in continuous cultures of BHK cells. *Biotechnol Bioeng* **66:** 104-113

Datta P, Linhardt RJ, Sharfstein ST (2013) An 'Omics Approach Towards CHO Cell Engineering. *Biotechnology and Bioengineering* **110:** 1255-1271

Davie JR (2003) Inhibition of histone deacetylase activity by butyrate. *The Journal of nutrition* **133:** 2485S-2493S

De Carvalho DD, Sharma S, You JS, Su SF, Taberlay PC, Kelly TK, Yang X, Liang G, Jones PA (2012) DNA methylation screening identifies driver epigenetic events of cancer cell survival. *Cancer cell* **21:** 655-667

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

Derouazi M, Martinet D, Besuchet Schmutz N, Flaction R, Wicht M, Bertschinger M, Hacker DL, Beckmann JS, Wurm FM (2006) Genetic characterization of CHO production host DG44 and derivative recombinant cell lines. *Biochemical and biophysical research communications* **340:** 1069-1077

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigo R (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research* **22:** 1775-1789

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloe D, Le Gall C, Schaeffer B, Le Crom S, Guedj M, Jaffrezic F (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics* **14:** 671-683

Dong Z, Zuber C, Pierce M, Stanley P, Roth J (2014) Reduction in Golgi apparatus dimension in the absence of a residential protein, N-acetylglucosaminyltransferase V. *Histochemistry and cell biology* **141:** 153-164

Doolan P, Clarke C, Kinsella P, Breen L, Meleady P, Leonard M, Zhang L, Clynes M, Aherne S, Barron N (2013) Transcriptomic analysis of clonal growth rate variation during CHO cell line development. *Journal of biotechnology*

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:** 592-608

Druz A, Chu C, Majors B, Santuary R, Betenbaugh M, Shiloach J (2011) A novel microRNA mmu-miR-466h affects apoptosis regulation in mammalian cells. *Biotechnol Bioeng* **108:** 1651-1661

Druz A, Son YJ, Betenbaugh M, Shiloach J (2013) Stable inhibition of mmu-miR-466h-5p improves apoptosis resistance and protein production in CHO cells. *Metabolic engineering* **16:** 87-94

Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung WK, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol I, Docking TR, Ho IY, Rokhsar DS, Chikhi R, Lavenier D, Chapuis G, Naquin D, Maillet N, Schatz MC, Kelley DR, Phillippy AM, Koren S, Yang SP, Wu W, Chou WC, Srivastava A, Shaw TI, Ruby JG, Skewes-Cox P, Betegon M, Dimon MT, Solovyev V, Seledtsov I, Kosarev P, Vorobyev D, Ramirez-Gonzalez R, Leggett R, MacLean D, Xia F, Luo R, Li Z, Xie Y, Liu B, Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Yin S, Sharpe T, Hall G, Kersey PJ, Durbin R, Jackman SD, Chapman JA, Huang X, DeRisi JL, Caccamo M, Li Y, Jaffe DB, Green RE, Haussler D, Korf I, Paten B (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome research* **21:** 2224-2241

Ernst W, Trummer E, Mead J, Bessant C, Strelec H, Katinger H, Hesse F (2006) Evaluation of a genomics platform for cross-species transcriptome analysis of recombinant CHO cells. *Biotechnology journal* **1:** 639-650

Fang Z, Martin J, Wang Z (2012) Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. *Cell & bioscience* **2:** 26

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. *Biotechnology and Bioengineering* **69:** 204-212

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:** 877-892

Flickinger MC, Goebel NK, Bibila T, Boyce-Jacino S (1992) Evidence for posttranscriptional stimulation of monoclonal antibody secretion by L-glutamine during slow hybridoma growth. *Journal of biotechnology* **22:** 201-226

Fogolin 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:** 179-191

Fraga MF, Ballestar E, Villar-Garea A, Boix-Chornet M, Espada J, Schotta G, Bonaldi T, Haydon C, Ropero S, Petrie K, Iyer NG, Perez-Rosado A, Calvo E, Lopez JA, Cano A, Calasanz MJ, Colomer D, Piris MA, Ahn N, Imhof A, Caldas C, Jenuwein T, Esteller M (2005) Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nature genetics* **37:** 391-400

Freshney RI (2005) Cloning and Selection. In *Culture of Animal Cells*. John Wiley & Sons, Inc.

Fullgrabe J, Kavanagh E, Joseph B (2011) Histone onco-modifications. *Oncogene* **30:** 3391-3403

Gaj T, Gersbach CA, Barbas CF, 3rd (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in biotechnology* **31:** 397-405

Gammell P, Barron N, Kumar N, Clynes M (2007) Initial identification of low temperature and culture stage induction of miRNA expression in suspension CHO-K1 cells. *Journal of biotechnology* **130:** 213-218

Gerstl MP, Hackl M, Graf AB, Borth N, Grillari J (2013) Prediction of transcribed PIWI-interacting RNAs from CHO RNAseq data. *Journal of biotechnology* **166:** 51-57

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. *Biotechnol Bioeng* **28:** 1376-1389

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* **108:** 1513-1518

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29:** 644-652

Gramer MJ (2014) Product quality considerations for mammalian cell culture process development and manufacturing. *Advances in biochemical engineering/biotechnology* **139:** 123-166

Greilhuber J, Volleth M, Loidl J (1983) Genome size of man and animals relative to the plant Allium cepa. *Canadian journal of genetics and cytology Journal canadien de genetique et de cytologie* **25:** 554-560

Grillari J, Fortschegger K, Grabherr RM, Hohenwarter O, Kunert R, Katinger H (2001) Analysis of alterations in gene expression after amplification of recombinant genes in CHO cells. *Journal of biotechnology* **87:** 59-65

Ha TK, Lee GM (2014) Effect of glutamine substitution by TCA cycle intermediates on the production and sialylation of Fc-fusion protein in Chinese hamster ovary cell culture. *Journal of biotechnology* **180:** 23-29

Hackl M, Jadhav V, Klanert G, Karbiener M, Scheideler M, Grillari J, Borth N (2014) Analysis of microRNA transcription and post-transcriptional processing by Dicer in the context of CHO cell proliferation. *Journal of biotechnology*

Hackl M, Jakobi T, Blom J, Doppmeier D, Brinkrolf K, Szczepanowski R, Bernhart SH, Honer Zu Siederdissen C, Bort JA, Wieser M, Kunert R, Jeffs S, Hofacker IL, Goesmann A, Puhler A, Borth N, Grillari J (2011) Next-generation sequencing of the Chinese hamster ovary microRNA transcriptome: Identification, annotation and profiling of microRNAs as targets for cellular engineering. *Journal of biotechnology* **153:** 62-75

Hammond S, Swanberg JC, Polson SW, Lee KH (2012) Profiling conserved microRNA expression in recombinant CHO cell lines using Illumina sequencing. *Biotechnol Bioeng* **109:** 1371-1375

Hernandez Bort JA, Hackl M, Hoflmayer H, Jadhav V, Harreither E, Kumar N, Ernst W, Grillari J, Borth N (2012) Dynamic mRNA and miRNA profiling of CHO-K1 suspension cell cultures. *Biotechnology journal* **7:** 500-515

Ho SC, Yang Y (2014) Identifying and engineering promoters for high level and sustainable therapeutic recombinant protein production in cultured mammalian cells. *Biotechnology letters*

Hossler P (2012) Protein glycosylation control in mammalian cell culture: past precedents and contemporary prospects. *Advances in biochemical engineering/biotechnology* **127:** 187-219

Hossler P, Khattak SF, Li ZJ (2009) Optimal and consistent protein glycosylation in mammalian cell culture. *Glycobiology* **19:** 936-949

Hossler P, Mulukutla BC, Hu WS (2007) Systems analysis of N-glycan processing in mammalian cells. *PloS one* **2:** e713

Hu W-S (2012) *Cell Culture Bioprocess Engineering*: http://www.cellprocessbook.com/, http://cbt.umn.edu/.

Hu WS, Dodge TC, Frame KK, Himes VB (1987) Effect of glucose on the cultivation of mammalian cells. *Developments in biological standardization* **66:** 279-290

Hussain H, Maldonado-Agurto R, Dickson AJ (2014) The endoplasmic reticulum and unfolded protein response in the control of mammalian recombinant protein production. *Biotechnology letters*

Jackman S, Raymond A, Birol I. (2013) Scaffolding large genomes using mate-pair sequencing and ABySS.

Jacob NM (2011) Development and application of genomic tools for process enhancement in Chinese hamster ovary cells. PhD Thesis, Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN

Jacob NM, Kantardjieff A, Yusufi FN, Retzel EF, Mulukutla BC, Chuah SH, Yap M, Hu WS (2010) Reaching the depth of the Chinese hamster ovary cell transcriptome. *Biotechnol Bioeng* **105:** 1002-1009

Jayapal KP, Goudar CT (2014) Transcriptomics as a tool for assessing the scalability of Mammalian cell perfusion systems. In *Mammalian Cell Cultures for Biologics Manufacturing*, Zhou W, Kantardjieff A (eds), Vol. 139, 2013/08/21 edn, pp 227-243. Springer-Verlag Berlin Heidelberg

Jayapal KP, Wlaschin KF, Hu W-S, Yap MGS (2007) Recombinant Protein Therapeutics from CHO Cells - 20 Years and Counting. *CHO Consortium: SBE Special Section***:** 40-47

Jefferis R (2009) Recombinant antibody therapeutics: the impact of glycosylation on mechanisms of action. *Trends in pharmacological sciences* **30:** 356-362

Jiang Z, Sharfstein ST (2008) Sodium butyrate stimulates monoclonal antibody over-expression in CHO cells by improving gene accessibility. *Biotechnol Bioeng* **100:** 189-194

Jimenez del Val I, Nagy JM, Kontoravdi C (2011) A dynamic mathematical model for monoclonal antibody N-linked glycosylation and nucleotide sugar donor transport within a maturing Golgi apparatus. *Biotechnology progress* **27:** 1730-1743

Johnson KC, Jacob NM, Nissom PM, Hackl M, Lee LH, Yap M, Hu WS (2011) Conserved microRNAs in Chinese hamster ovary cell lines. *Biotechnol Bioeng* **108:** 475-480

Johnson KC, Yongky A, Vishwanathan N, Jacob NM, Jayapal KP, Goudar CT, Karypis G, Hu WS (2013) Exploring the transcriptome space of a recombinant BHK cell line through next generation sequencing. *Biotechnol Bioeng*

Johnston RN, Beverley SM, Schimke RT (1983) Rapid spontaneous dihydrofolate reductase gene amplification shown by fluorescence-activated cell sorting. *Proc Natl Acad Sci U S A* **80:** 3711-3715

Jun SC, Kim MS, Baik JY, Hwang SO, Lee GM (2005) Selection strategies for the establishment of recombinant Chinese hamster ovary cell line with dihydrofolate reductase-mediated gene amplification. *Applied microbiology and biotechnology* **69:** 162-169

Kadouri A, Spier RE (1997) Some myths and messages concerning the batch and continuous culture of
animal cells. *Cytotechnology* **24:** 89-98

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research* **42:** D199-205

Kantardjieff A, Jacob NM, Yee JC, Epstein E, Kok YJ, Philp R, Betenbaugh M, Hu WS (2010a) Transcriptome and proteome analysis of Chinese hamster ovary cells under low temperature and butyrate treatment. *Journal of biotechnology* **145:** 143-159

Kantardjieff A, Nissom PM, Chuah SH, Yusufi F, Jacob NM, Mulukutla BC, Yap M, Hu WS (2009) Developing genomic platforms for Chinese hamster ovary cells. *Biotechnology advances* **27:** 1028-1035

Kantardjieff A, Seth G, McIvor S, Hu W-S (2010b) Genetic Manipulations of Mammalian Cells for Protein Expression. In *Manual of Industrial Microbiology and Biotechnology (3rd Edition)*
Baltz RH, Davies JE, Demain AL (eds), pp 330-344. American Society for Microbiology (ASM)

Kantardjieff A, Zhou W (2014) Mammalian cell cultures for biologics manufacturing. *Advances in biochemical engineering/biotechnology* **139:** 1-9

Kaufman RJ, Schimke RT (1981) Amplification and loss of dihydrofolate reductase genes in a Chinese hamster ovary cell line. *Molecular and Cellular Biology* **1:** 1069-1076

Kaufman RJ, Sharp PA (1982) Amplification and expression of sequences cotransfected with a modular dihydrofolate reductase complementary DNA gene. *Journal of Molecular Biology* **159:** 601-621

Kaufman RJ, Sharp PA, Latt SA (1983) Evolution of Chromosomal Regions Containing Transfected and Amplified Dihydrofolate Reductase Sequences. *Molecular and Cellular Biology* **3:** 699-711

Kaufman RJ, Wasley LC, Spillotes AJ, Gossels SD, Latt SA, Larsen GR, Kay RM (1985) Coamplification and coexpression of human tissue-type plasminogen activator and murine dihydrofolate reductase sequences in Chinese hamster ovary cells. *Molecular and Cellular Biology* **5:** 1750-1759

Kaufmann H, Fussenegger M (2003) Metabolic engineering of mammalian cells for higher protein yield. In *New Comprehensive Biochemistry*, Makrides SC (ed), Vol. Volume 38, pp 457-469. Elsevier

Kaufmann H, Mazur X, Fussenegger M, Bailey JE (1999) Influence of low temperature on productivity, proteome and protein phosphorylation of CHO cells. *Biotechnol Bioeng* **63:** 573-582

Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, Adachi J, Fukuda S, Aizawa K, Izawa M, Nishi K, Kiyosawa H, Kondo S, Yamanaka I, Saito T, Okazaki Y, Gojobori T, Bono H, Kasukawa T, Saito R, Kadota K, Matsuda H, Ashburner M, Batalov S, Casavant T, Fleischmann W, Gaasterland T, Gissi C, King B, Kochiwa H, Kuehl P, Lewis S, Matsuo Y, Nikaido I, Pesole G, Quackenbush J, Schriml LM, Staubli F, Suzuki R, Tomita M, Wagner L, Washio T, Sakai K, Okido T, Furuno M, Aono H, Baldarelli R, Barsh G, Blake J, Boffelli D, Bojunga N, Carninci P, de Bonaldo MF, Brownstein MJ, Bult C, Fletcher C, Fujita M, Gariboldi M, Gustincich S, Hill D, Hofmann M, Hume DA, Kamiya M, Lee NH, Lyons P, Marchionni L, Mashima J, Mazzarelli J, Mombaerts P, Nordone P, Ring B, Ringwald M, Rodriguez I, Sakamoto N, Sasaki H, Sato K, Schonbach C, Seya T, Shibata Y, Storch KF, Suzuki H, Toyo-oka K, Wang KH, Weitz C, Whittaker C, Wilming L, Wynshaw-Boris A, Yoshida K, Hasegawa Y, Kawaji H, Kohtsuki S, Hayashizaki Y (2001) Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685-690

Keeler KJ, Dray T, Penney JE, Gloor GB (1996) Gene targeting of a plasmid-borne sequence to a double-strand DNA break in Drosophila melanogaster. *Mol Cell Biol* **16:** 522-528

Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome research* **12:** 656-664

Kildegaard HF, Baycin-Hizal D, Lewis NE, Betenbaugh MJ (2013) The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology. *Current opinion in biotechnology* **24:** 1102-1107

Kim JY, Kim YG, Lee GM (2012) CHO cells in biotechnology for production of recombinant proteins: current state and further potential. *Applied microbiology and biotechnology* **93:** 917-930

Kim M, O'Callaghan PM, Droms KA, James DC (2011) A mechanistic understanding of production instability in CHO cell lines expressing recombinant monoclonal antibodies. *Biotechnol Bioeng*

Kim NS, Byun TH, Lee GM (2001) Key determinants in the occurrence of clonal variation in humanized antibody expression of cho cells during dihydrofolate reductase mediated gene amplification. *Biotechnology progress* **17:** 69-75

Kim NS, Kim SJ, Lee GM (1998a) Clonal Variability Within Dihydrofolate Reductase-Mediated Gene Amplified Chonese Hamster Ovary Cells: Stability in the Absence of Selective Pressure. *Biotechnology and Bioengineering* **60:** 679-688

Kim SJ, Kim N, Ryu CJ, Hong HJ, Lee GM (1998b) Characterization of Chimeric Antibody Producing CHO Cells in the Course of Dihydrofolate Reductase-Mediated Gene Amplification and Their Stability in the Absence of Selective Pressure. *Biotechnol Bioeng* **58:** 73-84

Kim SJ, Lee GM (1999) Cytogenetic Analysis of Chimeric Antibody-Producing CHO Cells in the Course of Dihydrofolate Reductase-Mediated Gene Amplification and Their Stability in the Absence of Selective Pressure. *Biotechnology and Bioengineering* **64:** 741-749

Klausing S, Kramer O, Noll T (2011) Bioreactor cultivation of CHO DP-12 cells under sodium butyrate treatment - comparative transcriptome analysis with CHO cDNA microarrays. *BMC proceedings* **5 Suppl 8:** P98

Korke R, Gatti Mde L, Lau AL, Lim JW, Seow TK, Chung MC, Hu WS (2004) Large scale gene expression profiling of metabolic shift of mammalian cells in culture. *Journal of biotechnology* **107:** 1-17

Krambeck FJ, Betenbaugh MJ (2005) A mathematical model of N-linked glycosylation. *Biotechnol Bioeng* **92:** 711-728

Kurano N, Leist C, Messi F, Gandor C, Kurano S, Fiechter A (1990) Growth kinetics of Chinese hamster ovary cells in a compact loop bioreactor. 3. Selection and characterization of an anchorage-independent subline and medium improvement. *Journal of biotechnology* **16:** 245-258

Lai T, Yang Y, Ng SK (2013) Advances in Mammalian cell line development technologies for recombinant protein production. *Pharmaceuticals (Basel)* **6:** 579-603

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10:** R25

Le H, Kabbur S, Pollastrini L, Sun Z, Mills K, Johnson K, Karypis G, Hu W-S (2012) Multivariate analysis of cell culture bioprocess data—Lactate consumption as process indicator. *Journal of biotechnology* **162:** 210-223

Le H, Vishwanathan N, Kantardjieff A, Doo I, Srienc M, Zheng X, Somia N, Hu WS (2013) Dynamic gene expression for metabolic engineering of mammalian cells in culture. *Metabolic engineering* **20:** 212-220

Lee S, Seo CH, Lim B, Yang JO, Oh J, Kim M, Lee B, Kang C (2011) Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic acids research* **39:** e9

Lee SM, Kim YG, Lee EG, Lee GM (2014) Digital mRNA profiling of N-glycosylation gene expression in recombinant Chinese hamster ovary cells treated with sodium butyrate. *Journal of biotechnology* **171:** 56-60

Leno M, Merten O-W, Hache J (1992) Kinetic studies of cellular metabolic activity, specific IgG production rate, IgG mRNA stability and accumalation during hybridoma batch culture. *Enzyme and Microbial Technology* **14:** 135-140

Lewis NE, Liu X, Li Y, Nagarajan H, Yerganian G, O'Brien E, Bordbar A, Roth AM, Rosenbloom J, Bian C, Xie M, Chen W, Li N, Baycin-Hizal D, Latif H, Forster J, Betenbaugh MJ, Famili I, Xu X, Wang J, Palsson BO (2013) Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome. *Nature biotechnology* **31:** 759-765

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26:** 589-595

Li J, Zhang C, Jostock T, Dubel S (2007) Analysis of IgG heavy chain to light chain ratio with mutant Encephalomyocarditis virus internal ribosome entry site. *Protein engineering, design & selection : PEDS* **20:** 491-496

Lim SF, Chuan KH, Liu S, Loh SO, Chung BY, Ong CC, Song Z (2006) RNAi suppression of Bax and Bak enhances viability in fed-batch cultures of CHO cells. *Metabolic engineering* **8:** 509-522

Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research* **38:** D346-354

Liu B, Spearman M, Doering J, Lattova E, Perreault H, Butler M (2014) The availability of glucose to CHO cells affects the intracellular lipid-linked oligosaccharide distribution, site occupancy and the N-glycosylation profile of a monoclonal antibody. *Journal of biotechnology* **170:** 17-27

Liu PQ, Chan EM, Cost GJ, Zhang L, Wang J, Miller JC, Guschin DY, Reik A, Holmes MC, Mott JE, Collingwood TN, Gregory PD (2010) Generation of a triple-gene knockout mammalian cell line using engineered zinc-finger nucleases. *Biotechnol Bioeng* **106:** 97-105

Liu Y, Schroder J, Schmidt B (2013) Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics* **29:** 308-315

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1:** 18

Ma C, Martin S, Trask B, Hamlin JL (1993) Sister chromatid fusion initiates amplification of the dihydrofolate reductase gene in Chinese hamster cells. *Genes & Development* **7:** 605-620

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM (2013) RNA-guided human genome engineering via Cas9. *Science* **339:** 823-826

Malphettes L, Freyvert Y, Chang J, Liu PQ, Chan E, Miller JC, Zhou Z, Nguyen T, Tsai C, Snowden AW, Collingwood TN, Gregory PD, Cost GJ (2010) Highly efficient deletion of FUT8 in CHO cell lines using zinc-finger nucleases yields cells that produce completely nonfucosylated antibodies. *Biotechnol Bioeng* **106:** 774-783

Mariani BD, Schimke RT (1983) Gene amplification in a single cell cycle in Chinese hamster ovary cells. *The Journal of Biological Chemistry* **259:** 1901-1910

Mariati, Yeo JH, Koh EY, Ho SC, Yang Y (2014) Insertion of core CpG island element into human CMV promoter for enhancing recombinant protein expression stability in CHO cells. *Biotechnology progress*

McAtee AG, Templeton N, Young JD (2014) Role of Chinese hamster ovary central carbon metabolism in controlling the quality of secreted biotherapeutic proteins. *Pharmaceutical Bioprocessing* **2:** 63-74

McLeod J, O'Callaghan PM, Pybus LP, Wilkinson SJ, Root T, Racher AJ, James DC (2011) An empirical modeling platform to evaluate the relative control discrete CHO cell

synthetic processes exert over recombinant monoclonal antibody production process titer. *Biotechnology and Bioengineering* **108:** 2193-2204

Meleady P, Doolan P, Henry M, Barron N, Keenan J, O'Sullivan F, Clarke C, Gammell P, Melville M, Leonard M, Clynes M (2011) Sustained productivity in recombinant Chinese Hamster Ovary (CHO) cell lines: proteome analysis of the molecular basis for a process-related phenotype. *BMC biotechnology* **11:** 78

Melville M, Doolan P, Mounts W, Barron N, Hann L, Leonard M, Clynes M, Charlebois T (2011) Development and characterization of a Chinese hamster ovary cell-specific oligonucleotide microarray. *Biotechnology letters* **33:** 1773-1779

Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M, Schorderet P, Bernstein BE, Jaenisch R, Lander ES, Meissner A (2008) Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454:** 49-55

Milbrandt JD, Heintz NH, White WC, Rothman SM, Hamlin JL (1981) Methotrexate-amplified Chinese Hamster Ovary cells have 135kb region amplified which contains dhfr gene. *Proc Natl Acad Sci U S A* **78:** 6043-6047

Miller JC, Tan S, Qiao G, Barlow KA, Wang J, Xia DF, Meng X, Paschon DE, Leung E, Hinkley SJ, Dulay GP, Hua KL, Ankoudinova I, Cost GJ, Urnov FD, Zhang HS, Holmes MC, Zhang L, Gregory PD, Rebar EJ (2011) A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **29:** 143-148

Minn AH (2005) Thioredoxin-Interacting Protein Is Stimulated by Glucose through a Carbohydrate Response Element and Induces beta Cell Apoptosis. *Endocrinology* **146:** 2397-2405

Moore A, Mercer J, Dutina G, Donahue CJ, Bauer KD, Mather JP, Etcheverry T, Ryll T (1997) Effects of temperature shift on cell cycle, apoptosis and nucleotide pools in CHO cell batch cultues. *Cytotechnology* **23:** 47-54

Mount DW (2007) Using the Basic Local Alignment Search Tool (BLAST). *CSH protocols* **2007:** pdb top17

Mulukutla BC (2012) Systems analysis of glucose metabolism of cultured mammalian cells. PhD Thesis, Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN

Mulukutla BC, Gramer M, Hu WS (2012) On metabolic shift to lactate consumption in fed-batch culture of mammalian cells. *Metabolic engineering* **14:** 138-149

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

Mussolino C, Morbitzer R, Lutge F, Dannemann N, Lahaye T, Cathomen T (2011) A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic acids research* **39:** 9283-9293

Nissom PM, Sanny A, Kok YJ, Hiang YT, Chuah SH, Shing TK, Lee YY, Wong KT, Hu WS, Sim MY, Philp R (2006) Transcriptome and proteome profiling to understanding the biology of high productivity CHO cells. *Molecular biotechnology* **34:** 125-140

Nivitchanyong T, Martinez A, Ishaque A, Murphy JE, Konstantinov K, Betenbaugh MJ, Thrift J (2007) Anti-apoptotic genes Aven and E1B-19K enhance performance of BHK cells engineered to express recombinant factor VIII in batch and low perfusion cell culture. *Biotechnology and Bioengineering* **98:** 825-841

Niwa H, Yamamura K-i, Miyazaki J-i (1991) Efficient selection for high expression transfectants using novel eukaryotic expression vector. *Gene* **108:** 193-200

Niwa R, Shoji-Hosaka E, Sakurada M, Shinkawa T, Uchida K, Nakamura K, Matsushima K, Ueda R, Hanai N, Shitara K (2004) Defucosylated Chimeric Anti-CC Chemokine Receptor 4 IgG1 with Enhanced Antibody-Dependent Cellular Cytotoxicity Shows Potent Therapeutic Activity to T-Cell Leukemia and Lymphoma. *Cancer Res* **64:** 2127-2133

Nordstrom KJ, Mirza MA, Larsson TP, Gloriam DE, Fredriksson R, Schioth HB (2006) Comprehensive comparisons of the current human, mouse, and rat RefSeq, Ensembl, EST, and FANTOM3 datasets: identification of new human genes with specific tissue expression profile. *Biochemical and biophysical research communications* **348:** 1063-1074

Nunberg JH, Kaufman RJ, Schimke RT, Urlaub G, Chasin LA (1978) Amplified dihydrofolate reductase genes are localized to a homogeneously staining region of a single chromosome in a methotrexate-resistant Chinese hamster ovary cell line. *Proc Natl Acad Sci U S A* **75:** 5553-5556

Nyberg GB, Balcarcel RR, Follstad BD, Stephanopoulos G, Wang DI (1999) Metabolic effects on recombinant interferon-gamma glycosylation in continuous culture of Chinese hamster ovary cells. *Biotechnol Bioeng* **62:** 336-347

O'Callaghan PM, McLeod J, Pybus LP, Lovelady CS, Wilkinson SJ, Racher AJ, Porter A, James DC (2010) Cell line-specific control of recombinant monoclonal antibody production by CHO cells. *Biotechnology and Bioengineering* **106:** 938-951

Oh SK, Vig P, Chua F, Teo WK, Yap MG (1993) Substantial overproduction of antibodies by applying osmotic pressure and sodium butyrate. *Biotechnol Bioeng* **42:** 601-610

Osterlehner A, Simmeth S, Gopfert U (2011) Promoter methylation and transgene copy numbers predict unstable protein production in recombinant Chinese hamster ovary cell lines. *Biotechnol Bioeng* **108:** 2670-2681

Ozturk SS, Riley MR, Palsson BO (1991) Effects of Ammonia and Lactate on Hybridoma Growth, Metabolism, and Antibody Production. *Biotechnol Bioeng* **39:** 418-431

Pendse GJ, Karkare S, Bailey JE (1992) Effect of Cloned Gene Dosage on Cell Growth and Hepatitis B Surface Antigen Synthesis and Secretion in Recombinant CHO cells. *Biotechnology and Bioengineering* **40:** 119-129

Prashad K, Mehra S (2014) Dynamics of unfolded protein response in recombinant CHO cells. *Cytotechnology*

Puck TT, Cieciura SJ, Robinson A (1958) Genetics of Somatic Mammalian Cell III. Long-Term Cultivation of Euploid Cells from Human and Animal Subjects. *Journal of Experimental Medicine***:** 945-959

Rodriguez J, Spearman M, Huzel N, Butler M (2005) Enhanced production of monomeric interferon-beta by CHO cells through the control of culture conditions. *Biotechnology progress* **21:** 22-30

Ronda C, Pedersen LE, Hansen HG, Kallehauge TB, Betenbaugh MJ, Nielsen AT, Faustrup Kildegaard H (2014) Accelerating genome editing in CHO cells using CRISPR Cas9 and CRISPy, a web-based target finding tool. *Biotechnol Bioeng*

Rupp O, Becker J, Brinkrolf K, Timmermann C, Borth N, Puhler A, Noll T, Goesmann A (2014) Construction of a Public CHO Cell Line Transcript Database Using Versatile Bioinformatics Analysis Pipelines. *PloS one* **9:** e85568

Ryll T, Valley U, Wagner R (1994) Biochemistry of growth inhibition by ammonium ions in mammalian cells. *Biotechnol Bioeng* **44:** 184-193

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marcais G, Pop M, Yorke JA (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome research* **22:** 557-567

Santiago Y, Chan E, Liu PQ, Orlando S, Zhang L, Urnov FD, Holmes MC, Guschin D, Waite A, Miller JC, Rebar EJ, Gregory PD, Klug A, Collingwood TN (2008) Targeted

gene knockout in mammalian cells by using engineered zinc-finger nucleases. *Proc Natl Acad Sci U S A* **105:** 5809-5814

Schlatter S, Stansfield SH, Dinnis DM, Racher AJ, Birch JR, James DC (2005) On the Optimal Ratio of Heavy to Light Chain Genes for Efficient Recombinant Antibody Production by CHO Cells. *Biotechnology progress* **21:** 122-133

Schmidt M, Schwarzwaelder K, Bartholomae C, Zaoui K, Ball C, Pilz I, Braun S, Glimm H, von Kalle C (2007) High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). *Nature methods* **4:** 1051-1057

Sealover NR, Davis AM, Brooks JK, George HJ, Kayser KJ, Lin N (2013) Engineering Chinese hamster ovary (CHO) cells for producing recombinant proteins with simple glycoforms by zinc-finger nuclease (ZFN)-mediated gene knockout of mannosyl (alpha-1,3-)-glycoprotein beta-1,2-N-acetylglucosaminyltransferase (Mgat1). *Journal of biotechnology* **167:** 24-32

Sekhavat A, Sun JM, Davie JR (2007) Competitive inhibition of histone deacetylase activity by trichostatin A and butyrate. *Biochemistry and cell biology = Biochimie et biologie cellulaire* **85:** 751-758

Seo JS, Min BS, Kim YJ, Cho JM, Baek E, Cho MS, Lee GM (2014) Effect of glucose feeding on the glycosylation quality of antibody produced by a human cell line, F2N78, in fed-batch culture. *Applied microbiology and biotechnology* **98:** 3509-3515

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:** 557-564

Seth G, Hossler P, Yee J, Hu W-S (2006a) Engineering Cells for Cell Culture Bioprocessing – Physiological Fundamentals
Cell Culture Engineering. Hu W-S (ed), Vol. 101, pp 119-164. Springer Berlin / Heidelberg

Seth G, Ozturk M, Hu WS (2006b) Reverting cholesterol auxotrophy of NS0 cells by altering epigenetic gene silencing. *Biotechnol Bioeng* **93:** 820-827

Seth G, Philp RJ, Denoya CD, McGrath K, Stutzman-Engwall KJ, Yap M, Hu WS (2005) Large-scale gene expression analysis of cholesterol dependence in NS0 cells. *Biotechnol Bioeng* **90:** 552-567

Seth G, Philp RJ, Lau A, Jiun KY, Yap M, Hu WS (2007b) Molecular portrait of high productivity in recombinant NS0 cells. *Biotechnol Bioeng* **97:** 933-951

Shen D, Kiehl TR, Khattak SF, Li ZJ, He A, Kayne PS, Patel V, Neuhaus IM, Sharfstein ST (2010) Transcriptomic responses to sodium chloride-induced osmotic stress: a study of industrial fed-batch CHO cell cultures. *Biotechnology progress* **26:** 1104-1115

Shen D, Sharfstein ST (2006) Genome-wide analysis of the transcriptional response of murine hybridomas to osmotic shock. *Biotechnol Bioeng* **93:** 132-145

Shields RL, Lai J, Keck R, O'Connell LY, Hong K, Meng YG, Weikert SH, Presta LG (2002) Lack of fucose on human IgG1 N-linked oligosaccharide improves binding to human Fcgamma RIII and antibody-dependent cellular toxicity. *J Biol Chem* **277:** 26733-26740

Shinkawa T, Nakamura K, Yamane N, Shoji-Hosaka E, Kanda Y, Sakurada M, Uchida K, Anazawa H, Satoh M, Yamasaki M, Hanai N, Shitara K (2003) The absence of fucose but not the presence of galactose or bisecting N-acetylglucosamine of human IgG1 complex-type oligosaccharides shows the critical role of enhancing antibody-dependent cellular cytotoxicity. *J Biol Chem* **278:** 3466-3473

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. *Genome research* **19:** 1117-1123

Soukas A, Socci ND, Saatkamp BD, Novelli S, Friedman JM (2001) Distinct Transcriptional Profiles of Adipogenesisin Vivo and in Vitro. *Journal of Biological Chemistry* **276:** 34167-34174

Steentoft C, Bennett EP, Schjoldager KT, Vakhrushev SY, Wandall HH, Clausen H (2014) Precision genome editing - a small revolution for glycobiology. *Glycobiology*

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (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:** 15545-15550

Takahashi K, Yamanaka S (2006) Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126:** 663-676

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **6:** 377-382

Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* **Chapter 4:** Unit 4 10

Tempel S (2012) Using and understanding RepeatMasker. *Methods Mol Biol* **859:** 29-51

Thaisuchat H, Baumann M, Pontiller J, Hesse F, Ernst W (2011) Identification of a novel temperature sensitive promoter in CHO cells. *BMC biotechnology* **11:** 51

Tigges M, Fussenegger M (2006) Xbp1-based engineering of secretory capacity enhances the productivity of Chinese hamster ovary cells. *Metabolic engineering* **8:** 264-272

Tijo JH, Puck TT (1958) Genetics of Somatic Mammalian Cells II. Chromosomal Constitution of Cells in Tissue Culture. *Journal of Experimental Medicine***:** 259-271

Trask BJ, Hamlin JL (1989) Early dihydrofolate reductase gene amplification events in CHO cells usually occur on the same chromosome arm as the original locus. *Genes & Development* **3:** 1913-1925

Trummer E, Ernst W, Hesse F, Schriebl K, Lattenmayer C, Kunert R, Vorauer-Uhl K, Katinger H, Muller D (2008) Transcriptional profiling of phenotypically different Epo-Fc expressing CHO clones by cross-species microarray analysis. *Biotechnology journal* **3:** 924-937

Trummer E, Fauland K, Seidinger S, Schriebl K, Lattenmayer C, Kunert R, Vorauer-Uhl K, Weik R, Borth N, Katinger H, Muller D (2006) Process parameter shifting: Part II. Biphasic cultivation-A tool for enhancing the volumetric productivity of batch processes using Epo-Fc expressing CHO cells. *Biotechnol Bioeng* **94:** 1045-1052

Tsai IJ, Otto TD, Berriman M (2010) Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome biology* **11:** R41

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:** 5116-5121

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

Underhill GH, George D, Bremer EG, Kansas GS (2003) Gene expression profiling reveals a highly specialized genetic program of plasma cells. *Blood* **101:** 4013-4021

Urlaub G, Chasin LA (1980) Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity. *Proc Natl Acad Sci U S A* **77:** 4216-4220

Urlaub G, Kas E, Carothers AM, Chasin LA (1983) Deletion of the Diploid Dihydrofolate Reductase Locus from Cultured Mammalian Cells. *Cell* **33:** 405-412

Urlaub G, Mitchell PJ, Kas E, Chasin LA, Funanage VL, Myoda TT, Hamlin J (1986) Effect of Gamma Rays at the Dihydrofolate Reductase Locus: Deletions and Inversions. *Somatic Cell and Molecular Genetics* **12:** 555-566

Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD, Holmes MC (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435:** 646-651

Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD (2010) Genome editing with engineered zinc finger nucleases. *Nature reviews Genetics* **11:** 636-646

Vallee C, Durocher Y, Henry O (2014) Exploiting the metabolism of PYC expressing HEK293 cells in fed-batch cultures. *Journal of biotechnology* **169:** 63-70

Vene´ R, Delfino L, Castellani P, Balza E, Bertolotti M, Sitia R, Rubartelli A (2010) Redox Remodeling Allows and Controls B-Cell Activation and Differentiation. *Antioxidants & Redox Signaling* **13:** 1145-1155

Vishwanathan N, Le H, Jacob NM, Tsao YS, Ng SW, Loo B, Liu Z, Kantardjieff A, Hu WS (2013) Transcriptome dynamics of transgene amplification in Chinese hamster ovary cells. *Biotechnol Bioeng*

Wahrheit J, Niklas J, Heinzle E (2014) Metabolic control at the cytosol-mitochondria interface in different growth phases of CHO cells. *Metabolic engineering* **23:** 9-21

Wang Y, Crawford DR, Davies KJA (1996) adapt33, a Novel Oxidant-Inducible RNA from Hamster HA-1 Cells. *Archives of Biochemistry and Biophysics* **332:** 255-260

Wang Y, Davies KJ, Melendez JA, Crawford DR (2003) Characterization of adapt33, a stress-inducible riboregulator. *Gene expression* **11:** 85-94

Wang Z, Rong YP, Malone MH, Davis MC, Zhong F, Distelhorst CW (2005) Thioredoxin-interacting protein (txnip) is a glucocorticoid-regulated primary response gene involved in mediating glucocorticoid-induced apoptosis. *Oncogene* **25:** 1903-1913

Weber W, Fussenegger M (2007) Inducible product gene expression technology tailored to bioprocess engineering. *Current opinion in biotechnology*

Whitford WG (2006) Fed-Batch Mammalian Cell Culture in Bioproduction. *BioProcess International***:** 30-40

Wippermann A, Klausing S, Rupp O, Albaum SP, Buntemeyer H, Noll T, Hoffrogge R (2014) Establishment of a CpG island microarray for analyses of genome-wide DNA methylation in Chinese hamster ovary cells. *Applied microbiology and biotechnology* **98:** 579-589

Wirth M, Hauser H (2008) Genetic Engineering of Animal Cells. In *Biotechnology*, pp 662-744. Wiley-VCH Verlag GmbH

Wlaschin KF, Hu W-S (2006) Fedbatch Culture and Dynamic Nutrient Feeding. **101:** 43-74

Wlaschin KF, Hu W-S (2007a) Engineering cell metabolism for high-density cell culture via manipulation of sugar transport. *Journal of biotechnology* **131:** 168-176

Wlaschin KF, Hu WS (2007b) A scaffold for the Chinese hamster genome. *Biotechnol Bioeng* **98:** 429-439

Wlaschin KF, Nissom PM, Gatti Mde L, Ong PF, Arleen S, Tan KS, Rink A, Cham B, Wong K, Yap M, Hu WS (2005) EST sequencing for gene discovery in Chinese hamster ovary cells. *Biotechnol Bioeng* **91:** 592-606

Wong DC, Wong KT, Lee YY, Morin PN, Heng CK, Yap MG (2006) Transcriptional profiling of apoptotic pathways in batch and fed-batch CHO cell cultures. *Biotechnol Bioeng* **94:** 373-382

Wong NS, Wati L, Nissom PM, Feng HT, Lee MM, Yap MG (2010) An investigation of intracellular glycosylation activities in CHO cells: effects of nucleotide sugar precursor feeding. *Biotechnol Bioeng* **107:** 321-336

Worton RG, Ho CC, Duff C (1977) Chromosome stability in CHO cells. *Somatic cell genetics* **3:** 27-45

Wurm F (2013) CHO Quasispecies—Implications for Manufacturing Processes. *Processes* **1:** 296-311

Wurm FM (2004) Production of recombinant protein therapeutics in cultivated mammalian cells. *Nature biotechnology* **22:** 1393-1398

Wurm FM, Hacker D (2011) First CHO genome. *Nature biotechnology* **29:** 718-720

Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, Chen W, Xie M, Wang W, Hammond S, Andersen MR, Neff N, Passarelli B, Koh W, Fan HC, Wang J, Gui Y, Lee KH, Betenbaugh MJ, Quake SR, Famili I, Palsson BO (2011) The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nature biotechnology* **29:** 735-741

Yang M, Butler M (2002) Effects of Ammonia and Glucosamine on the Heterogeneity of Erythropoietin Glycoforms. *Biotechnology progress* **18:** 129-138

Yang Y, Mariati, Chusainow J, Yap MG (2010) DNA methylation contributes to loss in productivity of monoclonal antibody-producing CHO cell lines. *Journal of biotechnology* **147:** 180-185

Yee JC, de Leon Gatti M, Philp RJ, Yap M, Hu WS (2008) Genomic and proteomic exploration of CHO and hybridoma cells under sodium butyrate treatment. *Biotechnol Bioeng* **99:** 1186-1204

Yee JC, Gerdtzen ZP, Hu WS (2009) Comparative transcriptome analysis to unveil genes affecting recombinant protein productivity in mammalian cells. *Biotechnol Bioeng* **102:** 246-263

Yoon SK, Hong JK, Lee GM (2004) Effect of simultaneous application of stressful culture conditions on specific productivity and heterogeneity of erythropoietin in Chinese hamster ovary cells. *Biotechnology progress* **20:** 1293-1296

Yoon SK, Kim SH, Lee GM (2003a) Effect of low culture temperature on specific productivity and transcription level of anti-4-1BB antibody in recombinant Chinese hamster ovary cells. *Biotechnology progress* **19:** 1383-1386

Yoon SK, Song JY, Lee GM (2003b) Effect of low culture temperature on specific productivity, transcription level, and heterogeneity of erythropoietin in Chinese hamster ovary cells. *Biotechnol Bioeng* **82:** 289-298

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18:** 821-829

Zhang B, Pan X, Cobb GP, Anderson TA (2007) microRNAs as oncogenes and tumor suppressors. *Developmental biology* **302:** 1-12

Zhou M, Crawford Y, Ng D, Tung J, Pynn AFJ, Meier A, Yuk IH, Vijayasankaran N, Leach K, Joly J, Snedecor B, Shen A (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:** 27-34

Zhou W, Chen CC, Buckland B, Aunins J (1997) Fed-batch culture of recombinant NS0 myeloma cells with high monoclonal antibody production. *Biotechnol Bioeng* **55:** 783-792

Zhu J (2012) Mammalian cell protein expression for biopharmaceutical production. *Biotechnology advances* **30:** 1158-1170

Zou N, Ditty S, Li B, Lo S-C (2003) Random priming PCR strategy to amplify and clone trace amounts of DNA. *Biotechniques* **35:** 758-765

Appendix

Table 0-1: List of primers for qRT-PCR analysis

Gene symbol	Ensembl ID of mouse or mouse homolog	Left primer	Right primer
β–actin	ENSMUSG00000029580	GTCGTACCACTGGCATTGTG	AGGGCAACATAGCACAGCTT
hIgG HC	-	GGCTTCTATCCCAGCGACATC	GGCGTGGTCTTGTAGTTGTTCTC
hIgG LC	-	GGGCGTTATCCACCTTCCA	CGTGGTGTGCCTGCTGAATA
mDHFR	ENSMUSG00000021707	TCTGTTTACCAGGAAGCCATGA	AATTCCTGCATGATCCTTGTCA
mGLUT5	ENSMUSG00000028976	CCTACATGATCGGAGGCAGT	CCAAGCTCCTCCTCCTTCTT

Table 0-2: List of genes in the hyper productivity gene set

Gene Symbol	Description	Fold Change in Selection (S)	Fold Change in Amplification (A)	Fold Change in High Producers vs Low Producers (H)	Set I (S∩H)	Set II (A∩H)	Set III (Q∩H)
Apoptosis							
Zfand5	zinc finger, AN1-type domain 5 Gene	-1.34	1.32	1.23	Y	N	N
Cell cycle							
Xrcc5	X-ray repair complementing defective repair in Chinese hamster cells 5 Gene	-1.75	1.09	-1.42	Y	N	N
Nde1	nuclear distribution gene E homolog 1 (A nidulans) Gene	-1.66	-1.05	-2.50	Y	N	N
Energy metabolism							
Rpe	ribulose-5-phosphate-3-epimerase Gene	-1.96	-1.20	-1.36	Y	N	N
Mrps24	mitochondrial ribosomal protein S24 Gene	-1.64	-1.23	-1.59	Y	N	N
Gene expression							
Tdrd3	tudor domain containing 3 Gene	-2.80	1.20	-1.69	Y	N	N
Pax3	paired box gene 3 Gene	-2.27	-1.01	-1.46	Y	N	N
Ctbp1	C-terminal binding protein 1 Gene	-1.89	1.07	-1.69	Y	N	N
Zfp346	zinc finger protein 346 Gene	-1.51	-1.04	-1.28	Y	N	N
Ptrf	polymerase I and transcript release factor Gene	-1.00	1.26	-2.17	Y	N	N
E2f6	E2F transcription factor 6 Gene	1.06	1.05	-1.36	Y	N	Y
Snrpc	U1 small nuclear ribonucleoprotein C Gene	1.22	-1.02	1.15	Y	N	N
Wtap	Wilms' tumour 1-associating protein Gene	1.22	-1.04	-1.05	Y	N	N
Cstf2t	cleavage stimulation factor, 3' pre-RNA subunit 2, tau Gene	1.29	1.16	1.20	Y	N	N
Zfp600	zinc finger protein 600 Gene	1.32	1.74	2.11	N	Y	N
Papola	poly (A) polymerase alpha Gene	1.38	-1.43	1.83	Y	N	N
Utp11l	UTP11-like, U3 small nucleolar ribonucleoprotein, (yeast) Gene	1.43	-1.04	-1.09	Y	N	N
Prpf31	PRP31 pre-mRNA processing factor 31 homolog (yeast) Gene	1.70	1.45	1.79	Y	N	N
Slu7	SLU7 splicing factor homolog (S. cerevisiae) Gene	1.75	-1.11	1.20	Y	N	N
Glycosylation							
Stt3b	STT3, subunit of the oligosaccharyltransferase complex, homolog B (S. cerevisiae) Gene	-1.21	1.21	-1.12	Y	Y	N

Gene Symbol	Description	Fold Change in Selection (S)	Fold Change in Amplification (A)	Fold Change in High Producers vs Low Producers (H)	Set I (S∩H)	Set II (A∩H)	Set III (Q∩H)
Dpm1	dolichol-phosphate (beta-D) mannosyltransferase 1 Gene	1.31	-1.13	1.19	Y	N	N
Ion transport							
Pxk	PX domain containing serine/threonine kinase Gene	-1.95	-1.20	-1.69	Y	N	N
Atp11b	ATPase, class VI, type 11B Gene	1.26	-1.08	1.27	Y	N	N
Lipid metabolism							
Cyp20a1	cytochrome P450, family 20, subfamily A, polypeptide 1 Gene	-2.56	-1.32	-1.67	Y	N	N
GH511842.1	CCUF6931.b1 CCUF Peromyscus maniculatus bairdii BW:Nb Peromyscus maniculatus bairdii cDNA clone CCUF6931 5', mRNA sequence.	-1.95	-1.46	-1.22	Y	N	N
Dagla	diacylglycerol lipase, alpha Gene	1.74	1.29	1.81	Y	N	N
Unannotated or other functions							
FI847326.1	CH232-004M13.C7 CHORI-232 Microtus ochrogaster genomic clone CH232-004M13, genomic survey sequence.	-2.15	1.22	1.61	Y	Y	N
Twsg1	twisted gastrulation homolog 1 (Drosophila) Gene	-2.06	-1.07	-1.81	Y	N	N
Adam9	a disintegrin and metallopeptidase domain 9 (meltrin gamma) Gene	-1.93	1.19	-1.87	Y	N	N
LOC680531	similar to CG3880-PA (LOC680531), mRNA	-1.66	-1.13	-1.42	Y	N	N
D17Wsu104e	DNA segment, Chr 17, Wayne State University 104, expressed Gene	-1.65	1.25	-1.08	Y	Y	N
Pcolce2	procollagen C-endopeptidase enhancer 2 Gene	-1.57	-1.02	-1.34	Y	N	N
Nt5dc2	5'-nucleotidase domain containing 2 Gene	-1.55	-1.07	-1.21	Y	N	N
Fam3a	family with sequence similarity 3, member A Gene	-1.51	1.06	-1.63	Y	N	N
Smc6	structural maintenance of chromosomes 6 Gene	-1.40	1.06	-1.27	Y	N	N
Parp1	poly (ADP-ribose) polymerase family, member 1 Gene	-1.31	1.02	-1.06	Y	N	N
Jmjd4	jumonji domain containing 4 Gene	1.35	1.13	1.54	Y	N	N

Gene Symbol	Description	Fold Change in Selection (S)	Fold Change in Amplification (A)	Fold Change in High Producers vs Low Producers (H)	Set I (S∩H)	Set II (A∩H)	Set III (Q∩H)
AL824707.7	Mouse DNA sequence from clone RP23-95O1 on chromosome 4 Contains the 5' end of the gene for the likely ortholog of H. sapiens chromosome 9 open reading frame 138 (C9orf138), a novel gene, the Rraga gene for Ras-related GTP binding A, the gene for the lik	1.37	-1.09	1.40	Y	N	N
AB370295.1	Eukaryotic synthetic construct DHFR, hGM-CSF, DHFR genes for dihydrofolate reductase, human granulocyte-macrophage colony stimulating factor, dihydrofolate reductase, complete cds, cell_line: Chinese Hamster Ovary cell DR1000L-4N	1.39	-1.46	1.22	Y	Y	N
Gpatch3	G patch domain containing 3 Gene	1.48	1.16	1.40	Y	N	N
CS391352.1	Sequence 627 from Patent WO2006025879	1.52	-1.70	-1.50	Y	N	N
FQ075827.1	FQ075827 Rattus norvegicus brain Sprague-Dawley Rattus norvegicus cDNA clone TL0AAA70YN24 3', mRNA sequence.	1.62	1.40	-1.23	Y	N	N
GH524702.1	CCUG5603.b1 CCUG Peromyscus maniculatus bairdii BW:fetus Peromyscus maniculatus bairdii cDNA clone CCUG5603 5', mRNA sequence.	1.68	1.43	1.92	Y	N	N
2410001C21Rik	RIKEN cDNA 2410001C21 gene Gene	1.71	1.15	1.48	Y	N	N
AC016791.23	Mus musculus chromosome 19, clone RP23-187B17, complete sequence	1.73	-1.01	1.26	Y	N	N
Ythdf1	YTH domain family 1 Gene	1.80	-1.16	1.40	Y	N	N
AL611927.21	Mouse DNA sequence from clone RP23-445E20 on chromosome 4 Contains the Hkr3 gene for GLI-Kruppel family member HKR3, the Tas1r1 gene for taste receptor, type 1, member 1, the Klhl21 gene for kelch-like 21 (Drosophila)	1.91	1.56	2.49	N	Y	N
AC132224.3	Mus musculus BAC clone RP24-554P12 from 1, complete sequence	3.21	-1.27	2.17	N	Y	N
Protein Processing							
Hspd1	heat shock protein 1 (chaperonin) Gene	-1.43	-1.04	-1.21	Y	N	N
Ufm1	ubiquitin-fold modifier 1 Gene	1.12	-1.04	1.10	N	Y	N
Protein Secretion							
Stk25	serine/threonine kinase 25 (yeast) Gene	-1.44	1.06	-1.23	Y	N	N

Gene Symbol	Description	Fold Change in Selection (S)	Fold Change in Amplification (A)	Fold Change in High Producers vs Low Producers (H)	Set I (S∩H)	Set II (A∩H)	Set III (Q∩H)
Stx7	syntaxin 7 Gene	1.14	-1.38	1.18	Y	N	N
Scamp2	secretory carrier membrane protein 2 Gene	1.16	1.26	1.12	Y	N	N
Snx21	sorting nexin family member 21 Gene	1.27	-1.40	1.46	N	Y	N
Protein Synthesis							
Pin4	protein (peptidyl-prolyl cis/trans isomerase) NIMA-interacting, 4 (parvulin) Gene	1.32	1.08	1.14	Y	N	N
Redox Balance							
Abcb6	ATP-binding cassette, sub-family B (MDR/TAP), member 6 Gene	-1.90	1.10	-1.56	Y	N	N
Signaling							
Ing1	inhibitor of growth family, member 1 Gene	-1.81	1.02	-1.77	Y	N	N
Smurf1	SMAD specific E3 ubiquitin protein ligase 1 Gene	-1.50	1.90	-1.17	Y	Y	N
Ptpn14	protein tyrosine phosphatase, non-receptor type 14 Gene	1.25	1.40	-1.16	Y	N	N
Rac1	RAS-related C3 botulinum substrate 1 Gene	1.30	1.19	-1.02	N	Y	N
Traf6	TNF receptor-associated factor 6 Gene	1.54	1.01	1.24	Y	N	N
NM_008350.4	Mus musculus interleukin 11 (Il11), mRNA	1.56	1.19	1.74	Y	N	N
Zfp128	zinc finger protein 128 Gene	1.57	1.44	1.86	Y	N	N
Gps2	G protein pathway suppressor 2 Gene	1.79	1.03	1.51	N	Y	N
Arhgap29	Rho GTPase activating protein 29 Gene	1.90	-1.28	-1.46	Y	N	N

Table 0-3: List of samples used for transcriptome assembly, RNA-seq and microarray in Chapter 4.

Name	*Description (mRNA Source)*	*Total Output*	*RNA-Seq*	*Microarray*
	Recombinant CHO-K1 (IgG) Normalized library Exponential growth phase, suspension culture GsuI digestion of polyA	84.75 Mbp		
	Recombinant CHO-K1 (IgG) Normalized library Exponential growth phase, suspension culture GsuI digestion of polyA	57.39 Mbp		
rDG_2	Recombinant DG44 (IgG) Exponential growth phase, suspension culture Serum independent	2.57 Gbp		Y
rDG_H	Recombinant DG44 (EPO)- High Producer Exponential growth phase, suspension culture Serum independent Weak promoter, strong selection	3.94 Gbp	Y	
rDG_L	Recombinant DG44 (EPO)- Low Producer Exponential growth phase, suspension culture Serum independent Weak promoter, strong selection	4.29 Gbp	Y	
rDG_4,Aza	Recombinant DG44 (IgG)- High Producer Exponential growth phase, suspension culture Serum independent 5-azacytidine treatment	4.47 Gbp	Y	
rDG_3	Recombinant DG44 (GFP)- Unamplified cell line Suspension culture 0 mM MTX treatment Control for 50 mM MTX treatment	0.55 Gbp	Y	
rDG_3M	Recombinant DG44 (GFP)- Unamplified cell line Suspension culture 50 mM MTX treatment	0.61 Gbp	Y	

Name	*Description (mRNA Source)*	*Total Output*	*RNA-Seq*	*Microarray*
DG44	Parental DG44 Exponential growth phase, adherent culture Serum dependent	3.32 Gbp	Y	Y
Brain	Chinese Hamster Brain Tissue Brain mRNA from one late adolescent virgin female hamster	4.25 Gbp	Y	Y
Liver	Chinese Hamster Liver Tissue Liver mRNA from one late adolescent virgin female hamster	6.59 Gbp	Y	Y
rDX_1	Parental CHO cell line Untransfected	9.20 Gbp		
Ovary	Chinese hamster ovary tissue			Y
DXB11	Parental DXB11			Y
CHO-K1	Parental CHO-K1			Y
rDX_1	DXB11-derived recombinant IgG producer 1			Y
rDX_1M	DXB11-derived recombinant IgG producer 1 20 nM MTX treatment			Y
rDX_3	DXB11-derived recombinant IgG producer 3			Y
rDX_3M	DXB11-derived recombinant IgG producer 3 20 nM MTX treatment			Y
rDX_2	DXB11-derived recombinant DHFR			Y
rDX_2M	DXB11-derived recombinant DHFR 20 nM MTX treatment			Y
rDX_4	DXB11-derived recombinant IgG TNF-alpha fusion protein producer			Y
rDG_2, NaBu	Recombinant DG44 (IgG) Late phase, 2mM butyrate for 24 hr Serum independent			Y
rDG_2, Late	Recombinant DG44 (IgG) Late phase, suspension culture Serum independent			Y
rDG_1	Recombinant DG44 (IgG TNF-alpha fusion protein producer) Exponential growth phase, suspension culture Serum independent			Y