Genomic analysis and engineering of Chinese Hamster Ovary cells for improved therapeutic protein production

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

Sofie Alice O'Brien

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

Advisor: Professor Wei-Shou Hu

May 2020

Acknowledgements

First, I would like to express gratitude to my advisor, Professor Wei-Shou Hu, for his help and guidance throughout the years. I have learned so much from him and grown both as a scientist and as a person over my time in graduate school. I will always be thankful for the time and effort he put into my training, and I will carry his lessons with me for the rest of my career.

Much of my work was the result of collaborations with many other talented scientists. I would like to thank Dr. Gary Dunny, Dr. Aaron Barnes, Dr. Rebecca Erickson, and the rest of the Dunny lab for their support at the beginning of my PhD. I thank Dr. Michael Smanski and Dr. Nik Somia for their help with vector design and the transgene swapping project. I am also grateful to Dr. Juhi Ojha and Dr. Paul Wu for our collaboration on the development of an algorithm for integration site analysis. Thank you to Dr. Casim Sarkar, Dr. Samira Azarin, and Dr. Scott McIvor for serving on my thesis committee, and providing valuable feedback on my work.

I am especially thankful for the wonderful members of the Hu lab who have supported me during graduate school. Thank you, Dr. Arpan Bandyopadhyay, for introducing me to the lab and helping me get started. To Dr. Christopher Stach, thank you so much for being a great teacher in the lab, I learned a lot about vector design and cell engineering from our many discussions. Thank you to all of the other members of the Hu group, current and former, that I was privileged to know: Meghan McCann, Dr. Conor O'Brien, Jen One, Dr. Kevin Ortiz-Rivera, Zion Lee, Thu Phan, Hansol Kim, Min Lu, Yen-An Lu, Janani Narayan, Dr. David Chau, Dr. Tung Le, Dr. Dong Seong Cho, Dr. Kyoungho Lee, Dr. Hsu-Yuan Fu, Dr. Ravali Raju, Dr. Yonsil Park, and Dr. Liang Zhao. I am so grateful for the memories we made together at our group parties, going apple picking, and going to the lake.

Thank you to the undergraduates I had the privilege of working with throughout graduate school: Alicia Zhang, Marina Raabe, Clarissa Kraft, and Daniela Batres. Training all of you and watching you progress as scientists has been one of the most rewarding experiences during my PhD. I was able to accomplish much more with your help.

Special thanks to Victoria Roberts for advice on figure design – my publications have really benefited from beautiful figures created with her help.

I would also like to thank the many friends I made throughout my time at Minnesota. Thank you to Meghan McCann, Jen One, Jen Markou, Erica Connell, and Beth Harris for our girl's board game group and long discussions about careers and life. To Meghan McCann, Andrew Allman, and Breezy Wentz, thank you for being great friends and neighbors. To Jacob and Tiffany Held, thank you for being our first friends in Minnesota, I will miss our regular dinners. To Jen One, thank you for helping me get through TAing and being my nerd buddy. To Matt Quan, thank you for your positivity, puns, and our numerous get-togethers.

Finally, I would like to thank my family for their never-ending support during the past 5+ years. Thank you to my parents, Vadim and Anna Bluvstein, and my sisters Julia, Rachel, and Debbie for believing in me. To my husband Conor O'Brien, I cannot thank you enough for your love and support. Thank you for being there for me, both as a partner and as a lab mate. I would not have made it through this without you.

Dedication

To my parents, Vadim and Anna, and to my husband Conor.

Abstract

Protein biologics have transformed the field of medicine in recent years. These complex molecules are produced in living cells, primarily Chinese Hamster Ovary (CHO) cells. Due to the importance of these therapeutic proteins to disease treatment, it is essential to improve the efficiency of their production, both to promote the development of new therapies, and to bring down the cost of manufacture. One of the most important components of the production process is the development of a cell line for protein production. Many features of a cell line, such as cellular growth, metabolic behavior, and the integration site of the gene encoding the protein, influence the resulting productivity and quality of the protein produced. In this work, characteristics of desirable integration sites are investigated using genomics, transcriptomics, and epigenomics, and an exploration of cellular metabolic behavior was performed through kinetic modeling and cell engineering.

In order to develop a cell line for therapeutic protein production, DNA encoding the product gene is integrated into the host cell genome. After selection, single cell cloning is performed to isolate clonal cell lines that originated from a single cell. For cell lines generated using random integration of the product gene, the integration site of the transgene becomes a unique signature for each cell clone and can be used for verification of clonality. An algorithm for the determination of integration sites from next generation sequencing data was developed and was able to distinguish between cell lines derived from different parent clones.

In addition to clonality verification, knowledge of the integration site can be used for clone selection, as the integration site of a transgene can affect its expression as well as the stability of that expression over time. To study regions of the genome which may have inherent instability, a clonal cell line with high productivity of IgG was successively single cell cloned, and high and low producing clones were isolated. We found that a region of the genome containing an integration site repeatedly lost copies in low producing clones, suggesting that some genomic regions are liable to instability.

To expand this work, whole genome sequencing from 23 cell lines was utilized to identify common regions of the genome prone to structural variability. The regions deleted

in each cell line roughly clustered by host cell, showing that different cell lineages may have different vulnerable regions. Though overall a low percentage of the genome is vulnerable to deletions, certain regions were more prone to variability than others.

The influence of the genomic landscape on transgene expression was further evaluated by identifying the integration sites of a randomly integrated transgene delivered via lentivirus to the CHO genome. A correlation was found between high expression of the transgene and high transcriptional activity and accessibility of the region surrounding the integration site. Using destabilized GFP as a reporter gene, high producing clones were isolated with this method. Furthermore, the transgene was replaced with a new product gene, highlighting a streamlined method for cell line development.

Though the integration site can be a major contributor to cell line productivity, it is not sufficient to obtain a high producing cell line. Other traits, such as metabolic behavior, can have a large influence on the growth and protein production capabilities of the cell line. Through the use of a kinetic model and cell engineering of a protein producing CHO cell, we investigated both the robustness of cell metabolism during the growth stage, but also reduced waste metabolite production in late stage culture. This study highlights the difficulty of manipulating metabolism, but also the utility of model-based predictions towards cell line engineering.

All together, these studies highlight some of the technological advances that are pushing improvements in cell line development. Through the increased use of omics technologies, a better understanding of a “good” integration site is being obtained. This together with the recent improvements in genome engineering techniques is allowing for targeted integration of transgenes or host cell engineering cassettes into pre-selected locations. These technologies will continue to drive the development of next generation cell lines with high, stable expression of transgenes and ideal behavior for high density culture.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description
ATAC-seq	Assay for transposase accessible chromatin using sequencing
bp	base pairs
CHO	Chinese hamster ovary
CMV	Cytomegalovirus
dGFP	Destabilized green fluorescent protein
ELISA	Enzyme-linked immunosorbent assay
GFP	Green fluorescent protein
GOI	Gene of interest
GOT1	Glutamic-oxaloacetic transaminase 1
IgG	Immunoglobulin G
IRES	Internal ribosome entry site
kbp	kilobase pairs
LDH	Lactate dehydrogenase
mAb	Monoclonal antibody
MOI	Multiplicity of infection
PCR	Polymerase chain reaction
PDK	Pyruvate dehydrogenase kinase
RMCE	Recombinase-mediated cassette exchange
RPKM	Reads per kilobase per million reads mapped
TCA cycle	The citric acid cycle
TNFR	Tumor necrosis factor receptor
VCD	Viable cell density

1 Introduction

Protein biopharmaceuticals, or biologics, have transformed the treatment of complex diseases such as cancer and rheumatoid arthritis. These therapeutic molecules differ from traditional drugs such as penicillin or aspirin in that they are difficult to produce outside of cell-based systems, owing to their large size and the extensive protein modifications required for functionality after synthesis. Currently, the vast majority of therapeutic proteins are produced in Chinese Hamster Ovary (CHO) cells, in large part due to their ability to produce large quantities of product with human-like post translational modifications, and their demonstration as safe hosts from a regulatory standpoint.

In the traditional method for generating a protein producing cell line, a plasmid containing the gene of interest (GOI) is randomly integrated into the genome, and selection is used to isolate a pool of cells, with some methods using amplification to increase the copy number of the GOI. Single cell cloning is then performed to isolate a cell line that originated from a single cell, with a certain number of copies of the GOI integrated into specific regions within the genome. When selecting a clone, the integration site of the GOI can have a profound effect on the characteristics of the resulting cell line.

Once a cell line is obtained, the cells are grown in large tank bioreactors to produce the therapeutic protein. The growth and behavior of the cells, and the resulting productivity, are highly influenced by cell metabolism. As they are a continuous cell line, CHO cells exhibit a Warburg-type metabolism, where they process large amounts of glucose into lactate, as opposed to routing the sugar into the mitochondria for the production of energy. In the context of a bioreactor, this production of lactate and other waste metabolites is undesirable, however, it is a hallmark of rapidly growing cells and cannot be easily modified without potential consequences to cell growth. In the interest of developing better cell lines for protein production, engineering cell lines to change their inherent metabolic behavior is an active field of study.

1.1 Thesis organization

This work focuses on understanding and improving the development of cell lines for therapeutic protein production. This goal is addressed on two fronts: understanding what

makes a desirable integration site from the transcriptional activity, accessibility, and stability perspective, and addressing potential production bottlenecks by engineering cell metabolism. The thesis is arranged into eight chapters. Chapter 2 examines recent advances in cell culture bioprocessing, and how they may be applied to future therapies. Chapter 3 describes a method for rapid identification and analysis of integration sites of the GOI in protein producing cell lines, and how the integration site may be used for proof of clonality, i.e. that a cell line originated from a single cell. Chapter 4 continues this study of integration sites and clonality by examining variability in successively subcloned cell lines at the chromosomal and gene copy number levels. To extend our understanding of genomic variability among CHO cell lines, analysis of whole genome sequencing from 23 cell lines is examined to identify potentially unstable genomic regions in Chapter 5. Chapter 6 investigates the correlation between expression of a GOI and the transcriptional activity and accessibility of the region surrounding the integration site. A cell line integrated into a desirable location was obtained, and recombinase mediated cassette exchange was used to replace the transgene to produce a new protein. Chapter 7 transitions to the discussion on engineering cell metabolism. A kinetic model of central metabolism was expanded to include glutamine metabolism and used to understand the behavior of cells engineered to reduce waste product generation. Finally, Chapter 8 provides a summary of the key results presented, and a discussion on future directions for this work.

2 Cell culture bioprocessing - the road taken and the path forward

Reproduced from: O'Brien, S. A. and Hu, W. S. Cell culture bioprocessing - the road taken and the path forward. (Manuscript submitted)

2.1 Summary

Cell culture processes are used to produce the vast majority of protein therapeutics, valued at over US$180 billion per annum worldwide. For more than a decade now, these processes have become highly productive. To further enhance capital efficiency, there has been an increase in the adoption of disposable apparatus and continuous processing, as well as a greater exploration of in-line sensing and various -omic tools to enhance process controllability and product quality consistency. These feats in cell culture processing for protein biologics will help accelerate the bioprocess advancements for virus and cell therapy applications.

2.2 Introduction

Mammalian cells have tremendous biosynthetic potential in producing complex proteins requiring difficult post-translational modifications that cannot be performed by microbial cells. For thirty decades, we have been extremely successful in harnessing this synthetic potential and have fostered a very large biomanufacturing sector. In this review, we highlight a number of innovations and renovations in cell culture process technology that facilitated the continued success of mammalian cell-based protein therapeutics, from process technology and product quality management, to improved cell lines and the employment of -omic tools and systems approaches (Figure 1). As the next generation of products for cell and gene therapy emerge, these new analytical assays and systems approaches will help shape the manufacturing process for these products to become robust in productivity and product quality.

2.3 Cell Culture Process Technology

For nearly three decades most cell culture processes were practiced in fed-batch mode, in which concentrated nutrient feed is added during cell cultivation to extend the production period, allowing cell concentration to reach a high level and for the product to accumulate. For over a decade, IgG products have achieved titers in the 10 g/L range, a level unimaginable when antibody products began to take off in mid 1990s. Continuous cell culture, almost invariably coupled with cell retention using an internal or external cell settling device to increase cell concentration and productivity, was for many years relegated to products that are labile to degradation or are produced at very low levels. In the past ten years, there has been increased interest in smaller, disposable reactors and continuous culture, instead of using large stainless-steel based tanks for fed-batch culture.

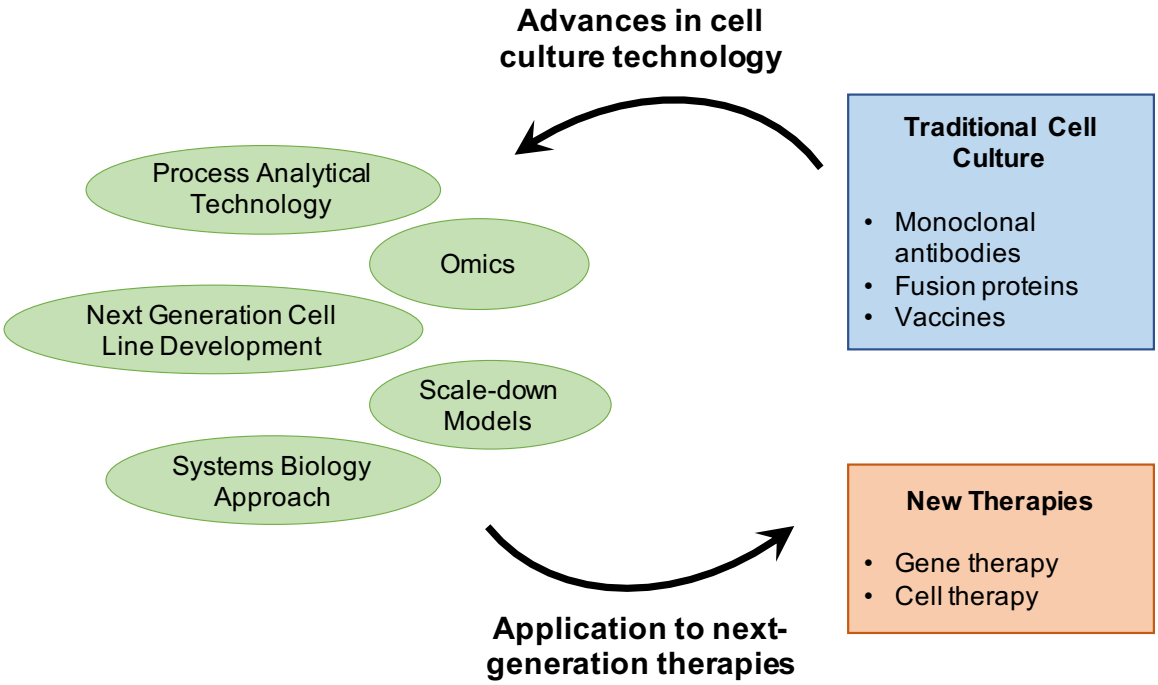

Figure 2-1. Major areas of advancement for cell culture bioprocessing.

The push toward continuous culture was driven by a need to increase plant throughput, due to the limits on the largest reactor size that can be constructed with plastic materials. Adoption of continuous operation was also facilitated by reduced media protein content and reduced membrane fouling, together with the success of cell retention hollowfiber devices with alternating tangential flow (ATF). Hollow fiber systems used for

cell retention are largely operated in laminar flow regions where a “tubular pinch” effect was predicted by fluid flow modeling, leading to larger, rigid spheres being lifted from the walls of the channel while submicron ones move to the membrane surface [1]. Whether tubular pinch occurs in the ATF device is not known as the direction of the axial flow is periodically reversed in the ATF system.

Studies have been performed to model the dynamics of fluid flow through ATF hollow fiber systems. The effect of cellular residence time in both the ATF device and the fluid transfer line on possible oxygen starvation and hydrodynamic damage to cells (via estimated energy dissipation) were evaluated [2]. A computational fluid dynamics (CFD) model was used to evaluate the effects of operating conditions on flux through an ATF hollow fiber and showed the presence of reverse flow across the membrane, a phenomenon known as Starling flow which is thought to reduce fouling [3].

A continuous culture can be operated at a steady state, thus keeping cell’s physiological state constant over time, and possibly delivering more consistent product quality. Most continuous cell culture processes are operated at a relatively slow growth rate compared to the maximal growth rate in a batch culture. At a low growth rate, a smaller cell purge rate can be used to maintain steady operation. Furthermore, a slower growth rate facilitates a switch to a low glycolytic flux state (for review see [4]). A strategy shown to improve perfusion cultures was to slow cell growth by reducing temperature or adding growth inhibitory valeric acid, reducing specific lactate production, increasing specific ammonia production, and increasing monoclonal antibody (mAb) productivity [5].

To maintain a constant cell concentration in a perfusion culture, some have adopted the classical turbidostat strategy, keeping the cell density constant by adjusting the dilution rate, feed nutrient concentration, or cell purge rate. An in-line capacitance probe that measures the viable cell concentration is often used for control [2]. Nonetheless, it is not unusual to see cell concentration and the various growth and metabolic indicators of specific rates fluctuate over a wide range. The complexity of allosteric regulations in cell metabolism allow for multiple metabolic states to exist for a given set of nutrient feed and dilution rate operating conditions. In such complex systems, the steady state that a system reaches is affected by the initial conditions and the culture trajectory. Manipulating the

startup culture conditions may lead to different steady states [6]. However, systematic discussion of perfusion culture dynamics and control strategies has not emerged.

2.4 Product Quality and Process Analytical Technology

Besides the capacitance sensor, Raman spectroscopy has become a commonly employed in-line sensor for monitoring various nutrient levels since its early application in cell culture processing [7]. The spectra acquired by an in-line Raman spectroscopic sensor are typically subjected to multivariate analysis or chemometrics to determine the concentration of the variable of interest using partial least squares. In the past few years, it has been adopted to monitor the concentrations of glucose, glutamate, glutamine, and ammonia. The concentration range of quantitative measurement is mostly in the ~1 mM range or higher. It has been shown that even with a relatively small number of training datasets, an online glucose measurement model could be established, and a PAT controller was developed to keep glucose at a low level (2g/L), simultaneously reducing the level of glycation (i.e. the covalent binding of a glucose molecule to an amino group in a protein) in the product protein [8].

In addition to glycation, the abundance level of other types of structural variants, including glycosylation, disulfide bond scrambling, and various proteolytic cleavages of the product protein, must also be controlled. Like glycation, disulfide bond cleavage and reorganization occur after the product protein molecules are secreted to the medium. This happens in the product recovery stage when the dissolved oxygen level is low, or in the presence of high levels of reductive agents. It was reported that cell lysis from shear damage in depth filtration caused the release of enzymes and NADPH, resulting in disulfide bond cleavage. This was mitigated by keeping the dissolved oxygen level high in the storage bag, thus decreasing the reducing environment in the clarified harvest [9]. Proteolytic cleavage by host cell enzymes may lead to loss of productivity and possible contamination of degradation fragments in the final product as seen in cell lines producing an IgG4-Fc fusion protein. Comparison of transcriptomes of cell lines with varying degree of proteolysis identified furin as the responsible protease, and subsequent use of a furin inhibitor in culture reduced the degradation of the fusion protein [10].

Advances in analytical technology are now enabling the characterization of N-glycans on glycosylated peptides of the erythropoietin-Fc fusion protein. A spatially distributed abundance level of sialic acid and fucose was observed [11]. Furthermore, the spatial distributions changed somewhat in different media and over different days of culture. The increased structural resolution of glycan heterogeneity will extend our understanding of their biological and clinical implications and enhance our capability to control them within an acceptable range. Spatial glycan distributions also attest to the complexity of manipulating glycosylation profiles, through either cell engineering or control of culture conditions. In this same study, o-glycan composition also changed over time and in different media [11]. The small number of o-glycans found on EPO-Fc was consistent with the limited o-glycosylation network predicted in CHO cells [12].

The level of host cell proteins (HCP) in the product must also be controlled to an acceptable regulatory level. Some enzymatic activity from residual HCPs may alter product characteristics over long-term storage. Proteomics were employed to identify contaminating HCPs, facilitating their chromatographic removal from the product [13] or knockout of the responsible gene to alleviate the problem [14]. Besides HCPs, endogenous retroviruses (ERVs) which are present in the genome of commonly used host cell lines are of safety and regulatory concern in biomanufacturing. ERVs have cryptic potential to generate infectious viral particles. They may also translocate, activate, or inactivate host cell genes and alter the host cell. Through the identification of integration loci of ERVs in Vero cells, it was shown that the likelihood of retro-translocation after the cell line is established is very low [15].

2.5 Cell lines

Cell lines for therapeutic protein production were traditionally constructed by random integration of linearized plasmid DNA containing a gene of interest (GOI) into the host cell genome, followed by selection and amplification of GOI copy number to ensure a high transcript level. Using this approach generates concatemers of the plasmid DNA, either in its entirety or fragments of it, and not all copies of the GOI are active. More recent approaches preserve the integrity of the GOI sequence when inserting it into the host cell genome at random sites or at a chosen locus. Random integration of the GOI in its entirety

can be done using transposase systems. Leap-in Transposase has been used to create higher producing pools with lower copy number than random plasmid integration [16], while piggyBac, Tol2, and Sleeping Beauty transposons were used to generate pools and clones with much higher volumetric productivity of TNFR-Fc than those made using plasmid only transfection [17].

Targeted integration approaches direct a GOI to a structurally stable, transcriptionally active location of the genome, with the aim of increasing the throughput and consistency of cell line development. A lentiviral vector was employed to integrate destabilized GFP (dGFP) and IgG along with sequence tags for recombinase mediated cassette exchange (RMCE) into the genome of CHO cells. After identifying a high producing clone with a single copy of the GOI through flowcytometric sorting and establishing the cell line, the RMCE site can be used to swap in a new target gene [18]. Another study utilized random integration to isolate a high producing, stable antibody producing clone, and then used RMCE to remove the antibody gene, creating a host cell line for targeted insertion of a new GOI [19]. CRISPR/Cas9 was later used to generate a landing pad for RMCE in this same site, creating a stable, high producing cell line [20]. In another case, multiple copies of GOI were inserted into a targeted integration host cell line to increase productivity [21].

A production cell line must give consistent productivity, growth characteristics, and product quality over a product's life cycle. Lacking structural and physiological understanding of these complex traits, the stability of a production cell line has been evaluated empirically, focusing on easily observable phenomena such as productivity and structural chromosomal stability. CHO cells are aneuploid, having a wide range of chromosome number, ranging from fewer than 20 to over 60. A small number of those chromosomes appear normal microscopically, and the rest are aberrant or consist of fused fragments. The karyotypes of CHO cells are inherently variable among different cell lines, and change over time as cells undergo replication [22, 23]. This appears to be different from a green monkey kidney cell line, Vero, which is frequently used in viral vaccine production [24]. Whether such karyotypic instability contributes to instability in productivity is not known. Another study showed that the distribution of chromosome

number was quickly reestablished after repeated single cell cloning, and one copy of the integrated GOI was repeatedly lost, thus suggesting structural instability in certain genomic regions [25].

With the advances in genome engineering, a variety of newer tools have been applied to engineer host cells. CRISPR interference (CRISPRi) utilizing dead Cas9 has been used to knockdown several genes associated with apoptosis (Bak, Bax, & Casp3) in CHO cells, decreasing caspase activity and increasing viable cell density [26]. CRISPR/Cas9 was used to knockout PKM1, resulting in reduced lactogenic behavior in late stage culture [27]. Overexpression of enzymes in phenylalanine-tyrosine catabolism or knockout of BCAT1 in the branched chain amino acid catabolic pathway reduced the accumulation of inhibitory byproducts in fed-batch culture [28]. Genes predicted to code for endogenous retroviruses in the CHO genome were knocked out, causing a reduction in retroviral RNA present in cell culture supernatant [29]. Host cell engineering will likely become more prevalent, especially with multi-gene manipulation to engineer pathways or traits to create a cell line with ideal metabolic behavior, production capabilities, and product quality characteristics.

Some proteins are difficult to express at high levels for various reasons: toxicity to the cells, complex post-translational modifications or quaternary structure, or instability after secretion, to name a few. Cell engineering and protein engineering are tools that can facilitate the production of such proteins. For example, the membrane proteins which are a potential immunogen for vaccination against respiratory syncytial virus have been engineered and converted to soluble, secreted molecules for production. A highly glycosylated trimer of HIV envelope protein has been produced in CHO cells and is being explored as a candidate HIV vaccine [30]. BMP-4 produced in CHO cells is actively re-internalized, thus contributing to its low productivity. Competitive inhibition of endocytosis by dextran sulfate increased its productivity [31]. Some difficult-to-express proteins, especially ones that are toxic to cells upon over-expression, such as ion channels, can be produced in a cell free system using CHO cell lysate, resulting in functional, properly folded proteins [32]. For such applications, CHO cells engineered to eliminate competing reactions and enhance the efficiency of expression will be in high demand.

2.6 Omics and Systems Approaches

In the past few years, genomic, transcriptomic, and some epigenomic assays have become widely applied to investigate different aspects of CHO cells in culture (reviewed in [33]). Sequence analysis of the mitochondrial genome of 22 CHO cell lines showed a large number of sequence variants, but most of them were cell line specific, including many in protein coding genes [34]. Not surprisingly, most variants are heteroplasmic with variant(s) existing in varying proportions within a cell line. Hence, even loss of function mutations are carried in some mitochondria within a cell line. These variants thus may or may not contribute to variability in cell lines or processes. Through transcriptome analysis of multiple high producing clones of two different products, a set of candidate genes associated with high productivity were identified [35]. Subsequent overexpression of various combinations of Erp27 and Erp57, both involved in protein folding and disulfide bond formation, and Foxa1, a transcription factor, resulted in increased productivity.

Besides CHO cell-based production of protein therapeutics, other cell culture processes, including virus production for vaccine and gene therapy applications, may benefit from -omic tools. Many cell lines used for vaccine production are of human origin, for which a well annotated genome is available. Two non-human cell lines frequently used in vaccine production, Vero cells and MDCK cells, were derived from African green monkey and dog respectively. The dog genome for many different breeds has long been available, and the genome of Vero cells has also been sequenced [24]. A 9 Mbp region of chromosome 12 was homozygously deleted in Vero cells, as compared to the reference genome of the African green monkey (from which Vero was originally derived). Among the genes lost was the type I interferon gene cluster, making the cells more susceptible to virus infections and more effective for virus production.

Increasingly, we shall be seeing multi-omic studies in cell bioprocessing. Using transcriptomic and metabolomic assays, the reduced supply of UDP-galactose was identified as the possible root cause of a changing glycosylation pattern in late stage culture, and this bottleneck was potentially due to a metabolic shift to a low glycolytic flux state. Supplementation of galactose in the late stage of culture increased the overall galactosylation level [36].

While the analysis of genomic and transcriptomic data requires bioinformatic tools, a model of the metabolic network is necessary to gain insight from metabolomic data. Most metabolic network models are based on stoichiometric balances. Since the system is invariably underdetermined, the calculated fluxes are dependent on the objective function selected and the solution algorithm used. Carbon isotope labeling is used to determine the split of carbon flux between key metabolic branch reactions and to constrain flux analysis solutions [37]. Major strides have been made in this area, and a potentially previously neglected reaction has been identified [38]. Additionally, a genome scale metabolic model has been developed for several CHO host cell lines, which can be used to better understand the effects of different bioprocess treatments and cell line engineering efforts [39]. It is worth noting that the metabolism of mammalian cells is compartmentalized. Without considering compartmentalization, the redox balance and even the carbon flow is skewed. And yet, the compartmentalization for some key reactions in amino acid metabolism and anaplerosis is still not fully understood. These models thus need to be updated as new knowledge emerges. Some have taken a kinetic perspective to model metabolism using a mechanistic kinetic model. A mechanistic kinetic model of glycolysis, the Krebs cycle, and the pentose phosphate pathway was used to identify combinations of gene expression alterations that can rewire glucose metabolism to a low flux state while meeting the constraints stipulated as requirements for growth [40].

A consequence of employing -omic assays is the generally increased parameter space for any problem related to cells. Multiplexing of cell culture experimentation to explore a wider region of parameter space is now routinely practiced in industrial bioprocess development. This type of equipment, for example the ambr system, can be automated for feeding and sampling, and is capable of pH control. These multiplex culture devices have been used to simulate perfusion culture by periodic gravity cell settling and medium exchange [41]. Multiplexing instrumentation is frequently used in scale-down studies that aim to simulate conditions in manufacturing reactors [42]. To harness the power of high throughput experimentation, a systems approach should be taken to integrate bioreactor operation, cell physiology, and growth. When the scale changes, many physical parameters related to the bioreactor, including aeration rate, mechanical stress, and mass

transfer rate, also change in different proportions. Changes to physical properties lead to changes in the chemical environment, which in turn alter the cell's physiology. These physiological changes further modify the chemical environment, forming a feedback loop. A systems model can integrate the physical effects, the chemical environment, and the physiological state of the cells to simulate cell growth, metabolic state, and the reactor environment. It can be a powerful tool for assessing process performance in different bioreactors scales and can assist in the design of experiments that capture critical parameters in scale translation.

2.7 Looking forward – conclusion

Protein therapeutics currently constitute the largest proportion of cell culture products. Nevertheless, it is worth noting that cell culture processing was rooted in viral vaccines. Adherent cells were adapted to suspension growth seven decades ago for the production of foot and mouth disease virus. The emergence of the SARS-CoV-2 virus has renewed the focus on vaccine technology. In addition to vaccines, the development of viruses as a gene delivery vehicle, and of immune and stem cells for therapeutic use, is accelerating. The current process for production of autologous cell therapies is individualized, as the treatment of a patient requires their own cells. The culture volume required for producing one dose of product is on the order of a few mL for vaccines, and about a liter for therapeutic cells and viral vectors. For personalized autologous cell products such as CAR-T cells, the production scale is thus rather small. Conversely, some vaccines and viral vectors are produced at a similar scale to protein biologics.

For personalized applications, automation to reduce manual steps has drawn development efforts [43]. For cell therapy, vaccine, and viral vector applications, the cell culture process is largely similar to that for the manufacturing of biologics, with emphasis on deploying disposable apparatus, achieving high cell concentration, and for cell therapy applications, keeping the product cells at a high viability and potency state. For these processes, the technologies developed for traditional biologics production are readily adoptable, including devices for high density continuous perfusion and the in-line capacitance and Raman spectroscopic sensors.

In the emerging cell and gene therapy areas, scant public information is available on the process, such as the kinetics of cell growth and product cell quality. For personalized medicine applications, there is a great need to understand the inherent differences in cells from different individuals which may cause their drastically different expansion and clinical potential. For gene vector production processes, one needs better control of the virus production process to minimize the proportion of viral particles which are defective or are void of the recombinant virus genome. Increasing process research efforts to fill these knowledge gaps will expedite the transition from clinical research and production to full-fledged manufacturing.

In closing, the success of protein biologics continues to drive the advancement of process technology, and these process advances in turn fuel the success of the industry. We see opportunities for process innovation in this mutually facilitating cycle. We also see the potential of various -omic technologies and systems designs enabling innovation. Many production processes have accumulated a trove of data, information, and knowledge over the life of the product. These processes are ripe for machine learning for further exploitation of process potential. It goes without saying that systems analysis and design relies on a system model. At the foundation of a system model is the heart of the process, i.e. cell metabolism, synthesis, and growth. Linking biological kinetics to reactor and process dynamics will allow for further process enhancement and provide a framework for developing advanced manufacturing processes for emerging cell technologies.

3 Multiplexed clonality verification of cell lines for protein biologic production

Reproduced with permission from: O'Brien, S. A., Ojha, J., Wu, P., & Hu, W. S. (2020). Multiplexed clonality verification of cell lines for protein biologic production. *Biotechnol Prog*, e2978. doi:10.1002/btpr.2978

3.1 Summary

During the development of cell lines for therapeutic protein production, a vector harboring a product transgene is integrated into the genome. To ensure production stability and consistent product quality, single-cell cloning is then performed. Since cells derived from the same parental clone have the same transgene integration locus, the identity of the integration site can also be used to verify the clonality of a production cell line. In this study, we present a high-throughput pipeline for clonality verification through integration site analysis. Sequence capture of genomic fragments that contain both vector and host cell genome sequences was used followed by next-generation sequencing to sequence the relevant vector-genome junctions. A Python algorithm was then developed for integration site identification and validated using a cell line with known integration sites. Using this system, we identified the integration sites of the host vector for 31 clonal cell lines from five independent vector integration events while using one set of probes against common features of the host vector for transgene integration. Cell lines from the same lineage had common integration sites, and they were distinct from unrelated cell lines. The integration sites obtained for each clone as part of the analysis may also be used for clone selection, as the sites can have a profound effect on the transgene's transcript level and the stability of the resulting cell line. This method thus provides a rapid system for integration site identification and clonality verification.

3.2 Introduction

Chinese hamster ovary (CHO) cells are one of the most commonly used cell lines used to produce therapeutic proteins [44]. They acquire the capability to produce these proteins by the introduction of a vector carrying the product gene and integration into the genome. After transfection, the resulting cells are heterogeneous and have a wide range of productivity and other properties. Single cell cloning is performed, and high producing cell lines are isolated as candidates to be the production cell line. To ensure the consistency and quality of the products produced by the cells over time, regulatory agencies require the demonstration of clonality, i.e. that cells originated from a single transfected ancestor cell [45].

After transfection and entry into the nucleus, the vector integrates into the genome of the host cell randomly. One cell may have one or more integration events depending on the vector used and its dose. The chance of two integration events occurring on the same site, either on both alleles of the same chromosome or in the genome of two different cells, is extremely low. A characteristic of clonally derived cells is thus that all cells within the population should have the same integration site(s) of the vector on the genome. Identifying the integration site of the product gene and demonstrating that two sublines of cells have the same integration site can thus be taken as evidence that the two originated from the same ancestor. Furthermore, the genome context of the integration site can be explored to reveal information on epigenetic accessibility, transcriptional activity [18], and even stability of the region [25].

Several Polymerase Chain Reaction (PCR) based assays such as inverse PCR [46], splinkerette-PCR [47], linear amplification mediated PCR (LAM-PCR) [48], and targeted locus amplification (TLA) [49] have been developed to identify integration sites of transgenes. These methods can be applied to demonstrate clonality. LAM-PCR for integration site identification has been used to analyze and track clonal lineages of blood cells following lentiviral gene therapy [50, 51], while TLA has been used in CHO cells for identification of clones with the same integration sites after pool selection and single cell cloning [52].

In this study, we established a high-throughput method that can be used to support monoclonality in different production cell lines generated using random integration of plasmids. By designing one set of probes for the common features of the host vector used to introduce the gene of interest, selective capture was utilized to simultaneously sequence vector-containing DNA from dozens of clonal cell lines from multiple independent vector integration events with different product genes. This method also allowed for a relatively simple bioinformatic analysis to identify these integration sites through the use of a Python algorithm. Cell lines were separable based on their integration sites, and we were able to demonstrate the clonality of the original cell populations. This method thus provides a rapid and cost-efficient tool for clonality verification and integration site identification in product producing CHO cells.

3.3 Materials and Methods

3.3.1 Cell line construction

The monoclonal antibody producing cell line, SH-87, has been described previously [53, 54]. The organization of the tricistronic vector used for introducing the transgene is illustrated in Figure 3-1A.

Cell lines used for clonality analysis were all derived from the same CHO-K1 host cell line using the vector shown in Figure 3-1B with product gene(s) inserted between the illustrated promoters and terminators. After transfection with linearized plasmid by electroporation, cells were selected using Puromycin in 96 wells, and surviving clones were isolated after monitoring by a CloneSelect Imager (Molecular Devices, San Jose, CA). Subsequent subcloning was performed via limiting dilution.

3.3.2 Probe design

For integration site analysis of cell line SH-87, probes (120bp in length) were focused on the ends of the linearized vector (Figure 3-1A). The first 900bp and the last 400bp of the linearized vector had 9x coverage with tiling (such that each location of the vector was covered by 9 probes), with the remainder of the vector having 1x coverage.

For clonality analysis of platform cell lines, probes (120bp in length) were designed to capture the regions of the linearized vector that are common in all transfected cell lines

at 5x coverage with tiling (Figure 3-1B). Sequences specific to individual cell lines, including that of the product gene, were thus not included in the probe design. Features in the vector with multiple occurrences, including the promoter and terminator, had reduced probe coverage.

3.3.3 Experimental methods

An overview of the experimental methods is shown in Figure 3-1C. Genomic DNA (gDNA) from SH-87 was provided courtesy of Dr. Yuansheng Yang and Dr. Dong-Yup Lee from ASTAR, Bioprocessing Technology Institute, Singapore. gDNA was sheared to an average length of 500bp, and library preparation was performed using an Agilent SureSelectXT Reagent kit (Agilent # G9611A, Santa Clara, CA). DNA fragments containing vector sequence were captured using the Agilent SureSelect Enrichment system. Briefly, short biotinylated RNA probes were hybridized to the pooled gDNA library, and streptavidin conjugated beads were used to capture DNA fragments that had hybridized to the probes. Captured DNA was eluted, amplified, and sequenced on half a lane of Illumina MiSeq (Illumina, San Diego, CA) using 250bp paired end reads for a total of 12.7 million reads.

For all other cell lines, gDNA was extracted using a Qiagen Blood and Cell Culture DNA Max Kit (Qiagen #13362, Valencia, CA). 100ng of gDNA from each cell line was sheared using a Covaris E220 ultrasonicator (Covaris, Woburn, MA) to obtain fragments with an average length of 250bp. Further library preparation was performed using an Agilent SureSelect HS Reagent Kit (Agilent #G9702A, Santa Clara, CA). Libraries for all cell lines were pooled prior to sequence capture following the Agilent SureSelect system instructions. Captured DNA was eluted, amplified, and sequenced on one lane of Illumina MiSeq (Illumina, San Diego, CA) using 100bp paired end reads for a total of 25 million reads.

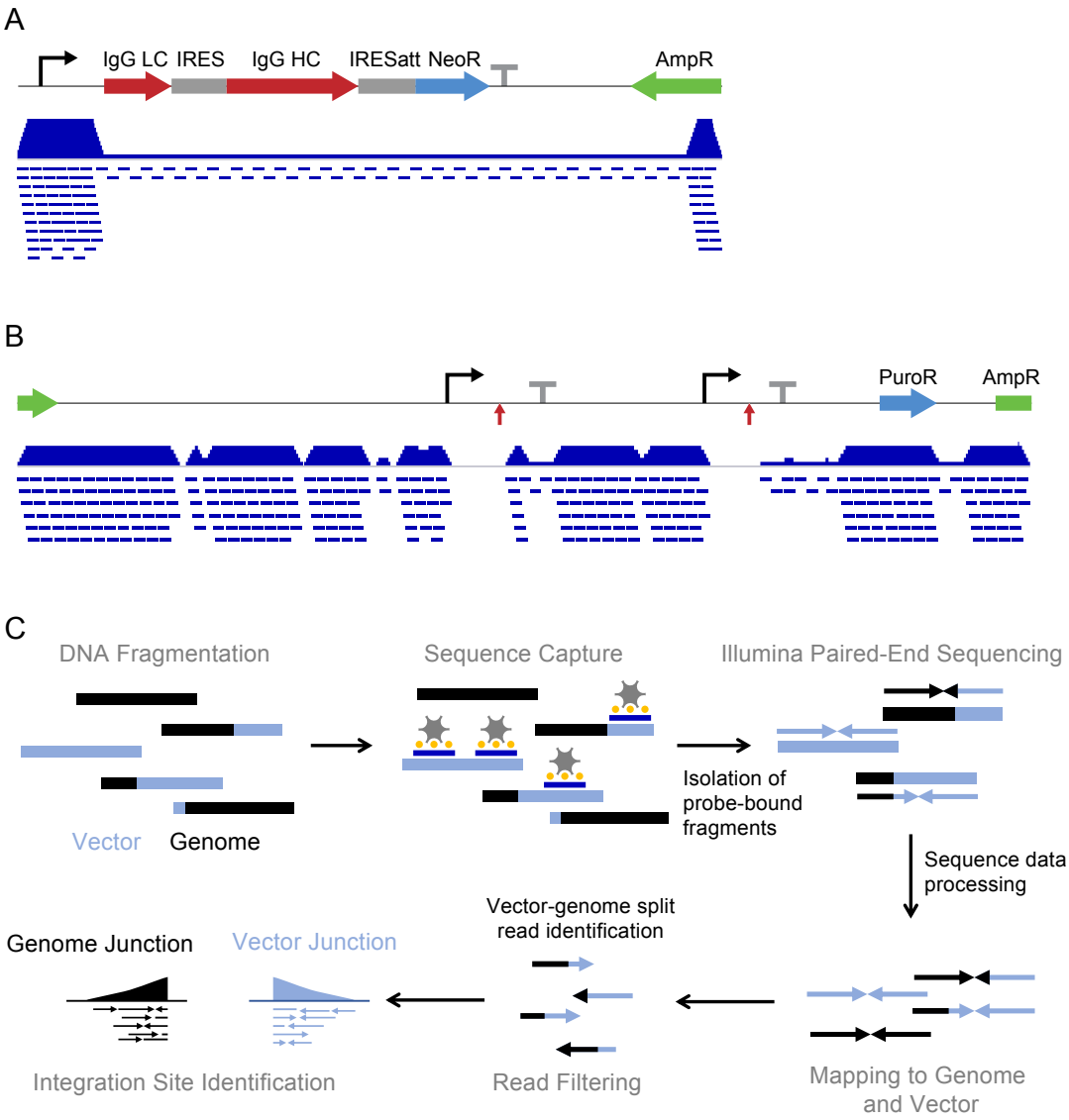

Figure 3-1. Overview of sequence capture method used. Probe coverage for SH-87 vector (A) and host vector from clonality analysis (B). The vector maps are shown at the top, with key features illustrated. The probe coverage at each location of the vector is shown below the vector map as a pileup ranging from zero to nine (A) or zero to five (B), and individual probe locations are denoted by short blue dashed lines below the pileup. Vector description for (A): A CMV (Cytomegalovirus) promoter (black solid angled arrow) drives expression of Immunoglobulin G Light Chain (IgG LC, red arrow), Immunoglobulin G Heavy Chain (IgG HC, red arrow), and Neomycin resistance (NeoR, blue arrow), all linked by IRES (Internal Ribosome Entry Site) elements (grey boxes), and followed by a SV40 early polyadenylation signal (grey solid t-shaped bar). The bacterial resistance marker (Ampicillin resistance, green arrow) is at the end of the vector next to the linearization site. Vector description for (B): The vector is linearized at the bacterial resistance marker (green arrow) and contains two sets of identical promoters (black solid angled arrows) and SV40 late polyadenylation signal terminators (grey solid t-shaped bars), as well as puromycin-N-acetyltransferase for selection (blue arrow). The product gene(s) are inserted between the promoters and terminators (red arrows). (C) Overview of experimental and bioinformatic methods used for integration site identification.

3.3.4 Data pre-processing

Fastq files from sequencing were trimmed to remove adapter sequences using Trimmomatic [55] (version 0.33). Reads were then mapped to the CriGri-PICR release of the Chinese hamster genome [56] using BWA-MEM [57] (BWA release 0.7.17). The host vector sequence was added to the genome sequence as an additional scaffold. Duplicate reads were removed using the MarkDuplicates command in Picard tools 2.18.16. The resulting SAM (Sequence Alignment/Map) file, which describes how reads are aligned to the genome, was used for further analysis. Samtools [58] (version 1.9) was used to create BAM files for visualization in IGV [59] (Integrative Genomics Viewer, version 2.4.19).

3.4 Results and Discussion

3.4.1 Development of an integration site analysis pipeline

An analysis pipeline was written in Python 3.6.3 to identify integration sites from mapped sequencing data (Figure 3-2A, algorithm is available at https://doi.org/10.13020/9wgm-mj51). Sequence capture data from cell line SH-87 was used for validation of the method, as integration sites from this cell line had been previously determined using whole genome sequencing and confirmed by PCR/Sanger sequencing [54].

The algorithm initially utilizes columns 3 and 4 (Scaffold and position) of the SAM file and the SA:Z tag added by BWA-MEM for chimeric alignments to identify reads which contain a split alignment between the vector and genome (Figure 3-2B). These alignments were then filtered for MAPQ (mapping quality score) > 30 and NM (number of mismatches) < 4. Reads mapping only to the genome, only to the vector, and unmapped reads were identified and counted for determining the number of on-target reads from sequence capture. For SH-87, there were 2.39×10^6 unique read pairs, and 18.4% of the read pairs mapped to the vector (see Table 3-1 for detailed mapping statistics). The low percentage of on-target reads was compensated by the high sequencing depth, providing a sufficient number of split-reads for integration site analysis.

Next, the CIGAR (Concise Idiosyncratic Gapped Alignment Report) tag for each alignment was used to find the exact vector-genome junction position. The CIGAR string

is a compact method to report how bases within a read align to the reference genome, specifying which bases match, are deleted/inserted, or are clipped (not aligned) in the case of split reads. Column 4 of the SAM file format reports the leftmost (smallest number) position that is aligned for the read alignment. If the CIGAR string begins with an alignment match (M), the length of the match is added to the position to obtain the vector-genome junction (Figure 3-2C). Otherwise, if the CIGAR string begins with bases that are not aligned (hard (H) or soft (S) clipping), the reported position is the first base that aligns, and thus is the location of the vector-genome junction (Figure 3-2D). Using this method, the position of the junction on the vector and genome for each read is identified, and the number of reads supporting each unique junction is tabulated. For paired end reads, if both ends of the pair support the junction, they were considered as only one count.

Table 3-1. Read mapping statistics for sequence capture data from SH-87. Each end of the paired end read is reported separately as R1 and R2.

	R1	**R2**
Number of Unique Reads (Library Size)	2,391,462	2,391,462
Reads mapping entirely to Vector	18.4%	18.4%
Reads mapping entirely to Genome	80.9%	80.0%
Split reads between vector and Genome	0.192%	0.175%
Low quality split reads	0.105%	0.099%
Unmapped reads	0.4%	1.4%

Table 3-2. Integration sites confirmed by algorithm for cell line SH-87. CHO-K1 Genome integration sites obtained from Yusufi, et al. 2016 [54].

Site	**CHO-K1 Genome Scaffold**	**Scaffold Position**	**CriGri-PICR Genome Scaffold**	**Scaffold Position**	**Vector Position**
A	NW_003613840.1	743320	NW_020822636.1	2762822	9
B	NW_003613840.1	747517	NW_020822636.1	2767055	505
D	NW_003613840.1	747631	NW_020822636.1	2767169	654
E	NW_003614673.1	136752	NW_020822636.1	898236	4466
C	NW_003614673.1	141082	NW_020822636.1	902565	1226
F	NW_003616992.1	80949	NW_020822636.1	177576	73

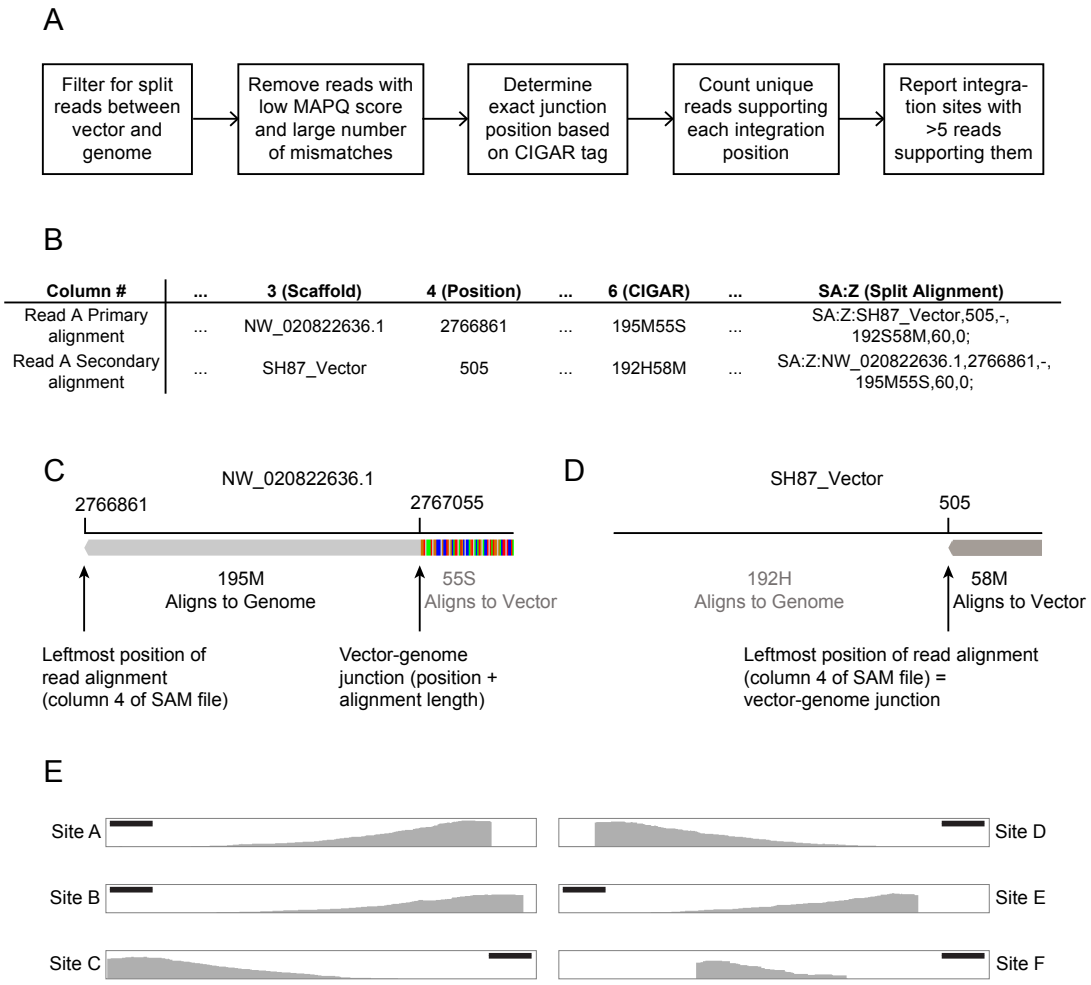

Figure 3-2. Integration site analysis pipeline and validation. (A) Procedure to identify integration sites. (B) SAM file alignments for a split read from mapped SH-87 sequence capture data. Primary and secondary alignments for the read are shown, along with select columns. (C) Illustration of primary alignment from (B) to Chinese Hamster genome. Vector-genome junction position is adjusted to account for portion of read aligned to genome. (D) Illustration of secondary alignment from (B) to SH-87 vector. Vector-genome junction does not need to be adjusted and is equal to the position listed in the SAM file. (E) IGV read pileups of identified integration sites for SH-87. Scale bar represents 50bp.

Due to the high sequencing depth of SH-87, a minimum of 15 reads or pairs of support were required for a vector-genome junction to be called an integration site. Previously, six vector-genome junctions from integration of the tricistronic vector in cell line SH-87 were identified by whole genome sequencing [54]. All six junctions were found by the algorithm (Table 3-2). To confirm these results, the genomic integration sites were visualized using IGV (Figure 3-2E). Each integration site has a sharp boundary at the vector-genome junction. Additionally, the read pileup depth decreases as the distance from

the integration site increases, as would be expected since the probability of a read being captured by vector probes diminishes with increasing distance from the junction. By using sequence capture and the Python algorithm, we were successfully able to determine the integration sites for SH-87 at a much lower sequencing and bioinformatic cost than whole genome sequencing.

3.4.2 Application of pipeline for confirmation of clonality

To apply this method to the confirmation of clonality, gDNA of 31 different cell lines derived from five independent clonal cell lines (denoted as cell lineages A-E) were used in this study. Cell lines characterized from lineages A, B, and E consisted of a clonal parental cell line and a set of subclones. Unrelated cell lines C-1 and D-1 were added to the analysis to increase the diversity of cell lines examined. Cell line A-1 was run in duplicate (labeled A-1 and A-1D) for a total of 32 samples. gDNA from these samples was sheared, and fragments containing probed vector regions were captured and sequenced.

After sequencing, the library size (number of unique read pairs) for each sample ranged from 0.46 - 5.6 x 10^5 (Figure 3-3A). The percent of on-target reads (reads which mapped either entirely or partially to the vector) was greater than 50% for all samples, with the majority of samples having >70% of their reads containing vector sequence (Figure 3-3B). This corresponded to an enrichment ratio (reads mapped to vector/reads not mapped to vector) of between 1.2 and 18x, depending on the sample.

The library size varied with cell lineage, with the cell lines from lineage E having on average significantly more unique reads after sequencing than those from lineage A or B (Figure 3-4A, t-test, $p<0.002$). Only one cell line was analyzed from each of cell lineages C and D, so these were not included in the comparison. The percent of on-target reads mapping to the vector also varied. Samples of cell lines from lineage E had a significantly higher percent of vector reads on average than lineages A or B (Figure 3-4B, t-test, $p<0.002$). There are several potential reasons for the bias in on-target read percentage and library size. gDNA from all cell lines was pooled at equal amounts after barcoding, so it is likely that inherent differences in the genome accounted for this difference. Cell lines from lineage E had three integration loci as opposed to the one locus found in populations from lineages A-D, and thus gDNA samples from lineage E would have higher vector sequence

content. Despite this variability, the sequencing depth for each sample was sufficient for integration site analysis.

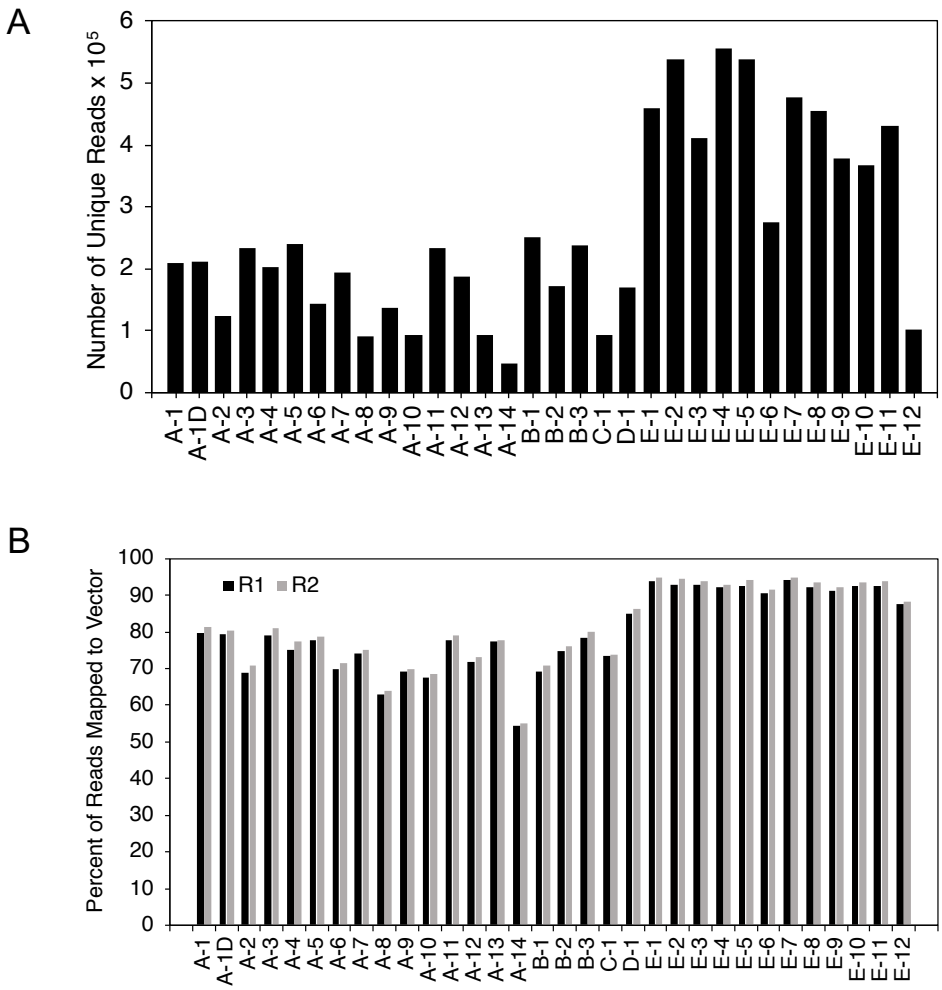

Figure 3-3. Sequencing statistics for cell lines used for clonality analysis. (A) Number of unique reads for each cell line after removal of PCR duplicates. (B) Percent of reads mapped to the vector for each cell line. Each end of the paired end read is reported separately as R1 and R2.

3.4.3 Cell lines from the same lineage have common integration sites

The results of the integration site analysis are shown in Figure 3-4C and Table 3-3. As the library size for these cell lines was lower, a minimum of five reads or pairs of support were required for a vector-genome junction to be called an integration site. Each integration site identified was on a different scaffold. Cells from lineage A had one integration locus, and both ends of the vector insertion were identified, approximately 30bp

apart. One end of a single integration locus was called by the algorithm for cells from lineages B, C, and D. This may be a result of complex sequence rearrangements, such as concatenations of the fragmented vector, at the other end of the vector-genome junction that would prevent proper mapping of the integration site. Incomplete mapping would also occur if the vector-genome junction was in a location that either was not probed or was not included in the vector sequence for mapping (such as the product gene), preventing the capture or mapping of DNA fragments from the integration site region. Three integration loci on different scaffolds were called for cells from lineage E, with both ends of the insertion found approximately 100bp apart for one of the three loci. Importantly, cell lines from the same lineage were found to have the same set of integration sites.

Table 3-3. Description of integration sites called by the algorithm for clonality analysis.

Site #	Cell Lines	Genome Location	Vector Junction
1a	All lineage A cell lines	Chromosome 1, exon, protein-coding	Between 2nd poly A and bacterial resistance
1b	All lineage A cell lines	Chromosome 1, exon, protein-coding	Between 1st poly A and 2nd promoter
1c	A-2	Chromosome 1, exon, predicted long non-coding RNA	Between 1st poly A and 2nd promoter
2	All lineage B cell lines	Chromosome 2, intron, protein-coding	Puromycin-N-acetyltransferase gene
3	C-1	Chromosome 6, intergenic region	Bacterial resistance on 3' end of linearized vector
4	D-1	Chromosome 5, intron, protein-coding	Between 2nd poly A and bacterial resistance
5	All lineage E cells lines	Chromosome 3, exon, protein-coding	Bacterial resistance on 5' end of linearized vector
6a	All lineage E cells lines	Chromosome 4, intergenic region	2nd poly A
6b	All lineage E cell lines except E-1	Chromosome 4, intergenic region	Bacterial resistance on 5' end of linearized vector
6c	E-2, E-10	Chromosome 4, intergenic region	1st promoter
7	All lineage E cells lines	Chromosome 4, exon, protein-coding	Bacterial resistance on 5' end of linearized vector

Several cell sublines (A-2, E-2, and E-10), had an additional vector-genome junction on the same scaffold as the integration site, either 1kbp (Site 1C for A-2, Figure 3-4C) or 4.7kbp (Site 6C for E-2 and E-10, Figure 3-4C) downstream of the main integration locus. As these additional vector-genome junctions were very close to the prevalent integration locus, it is highly unlikely that these are independent integration events. Rather, this could be the result of a genomic rearrangement or duplication in that scaffold that resulted in an additional integration junction, even though amplification was not performed during cell line development. Genomic heterogeneity has been previously shown after subcloning, with different subclones presenting a gain or loss of transgene copies [25].

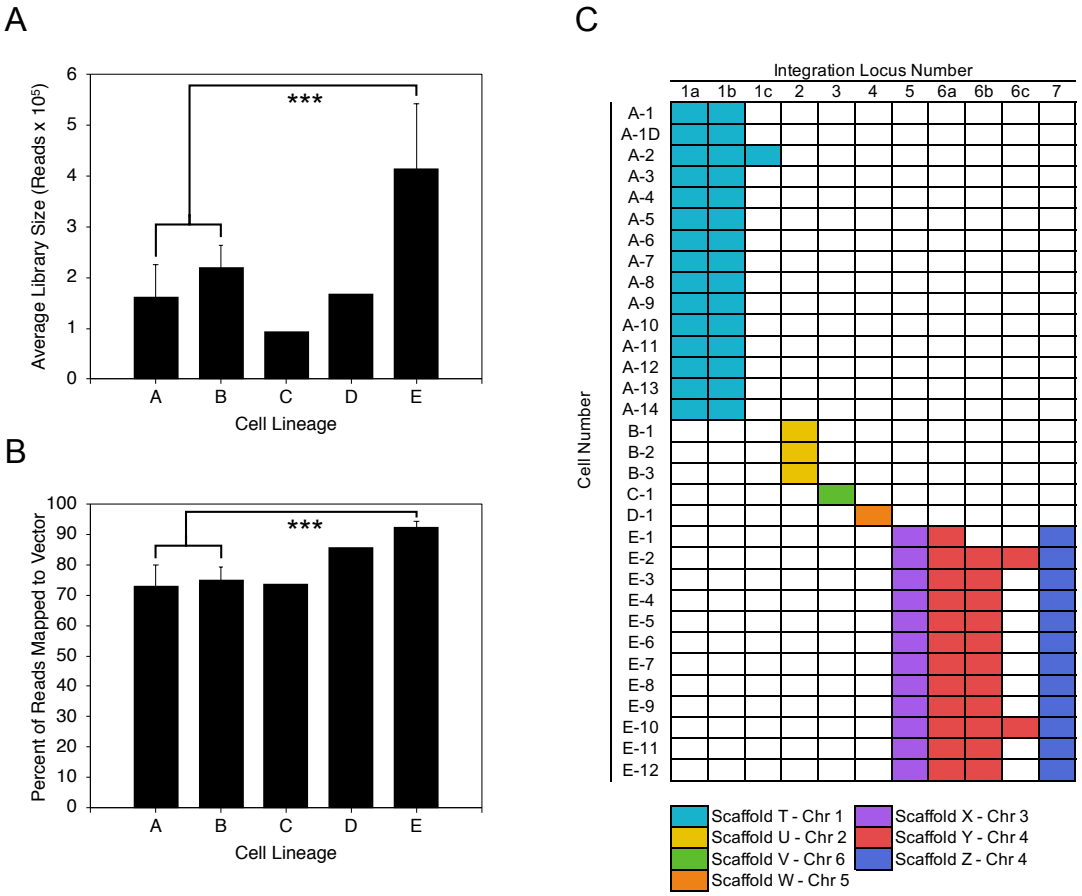

Figure 3-4. Results of integration site analysis algorithm. (A) Average number of unique reads for cell lines from each cell lineage (Number of cell lines for each lineage: A, n=14; B, n=3; C, n=1; D, n=1; E, n=14). *** $p < 0.002$, t-test. (B) Average percent of reads mapped to the vector for different cell lineages (*** $p < 0.002$, t-test). (C) Result of integration site identification. Each unique vector-genome junction is represented as an integration locus (numbered 1-7), and integration loci are colored by scaffold. The scaffolds and corresponding chromosomes are labeled below the chart.

Each integration site was visualized in IGV (Figure 3-5A-G). The vector-genome junctions are the same at the base by base level for all cell lines from the same lineage, with the read pileup depth decreasing as the distance from the integration site increases.

Only one side of the vector insertion was called by the algorithm for cell line C-1, but on the IGV pileup, both ends are visible, less than 50bp apart (Figure 3-5C). Upon further investigation, it was determined that the vector side of the split reads at this locus resides in the terminator element that is duplicated in the vector (Figure 3-1B). Since the vector part of these reads maps identically to two different locations on the vector, BWA was not able to assign an alignment, and these reads were only mapped to the genome. This however does not affect the capacity of the algorithm to verify clonality by integration site analysis; such a site would not map properly in any of the clones. Additionally, not every integration site needs to be identified to show clonality, as the likelihood that two independently derived cell lines would have even a single shared integration site is low. A longer sequencing read length may help avoid these types of non-unique mapping events in the future.

The algorithm did not find integration locus 6b in cell line E-1 (Figure 3-4C). This integration locus is present in this cell line, as can be seen in Figure 3-5F, on the right side of the read pileup. The read depth at the integration junction was below the set threshold of five reads, and so the algorithm did not qualify this location as a true integration site. Increased read depth for this sample would have given more confidence in this integration locus.

In this study, we used our method to verify that each subclone had originated from its clonal parent. This workflow can also be used to test clonality or detect non-clonal cells by sequencing a large number of subclones; the presence of subclones with non-consensus integration site(s) would suggest possible non-clonality.

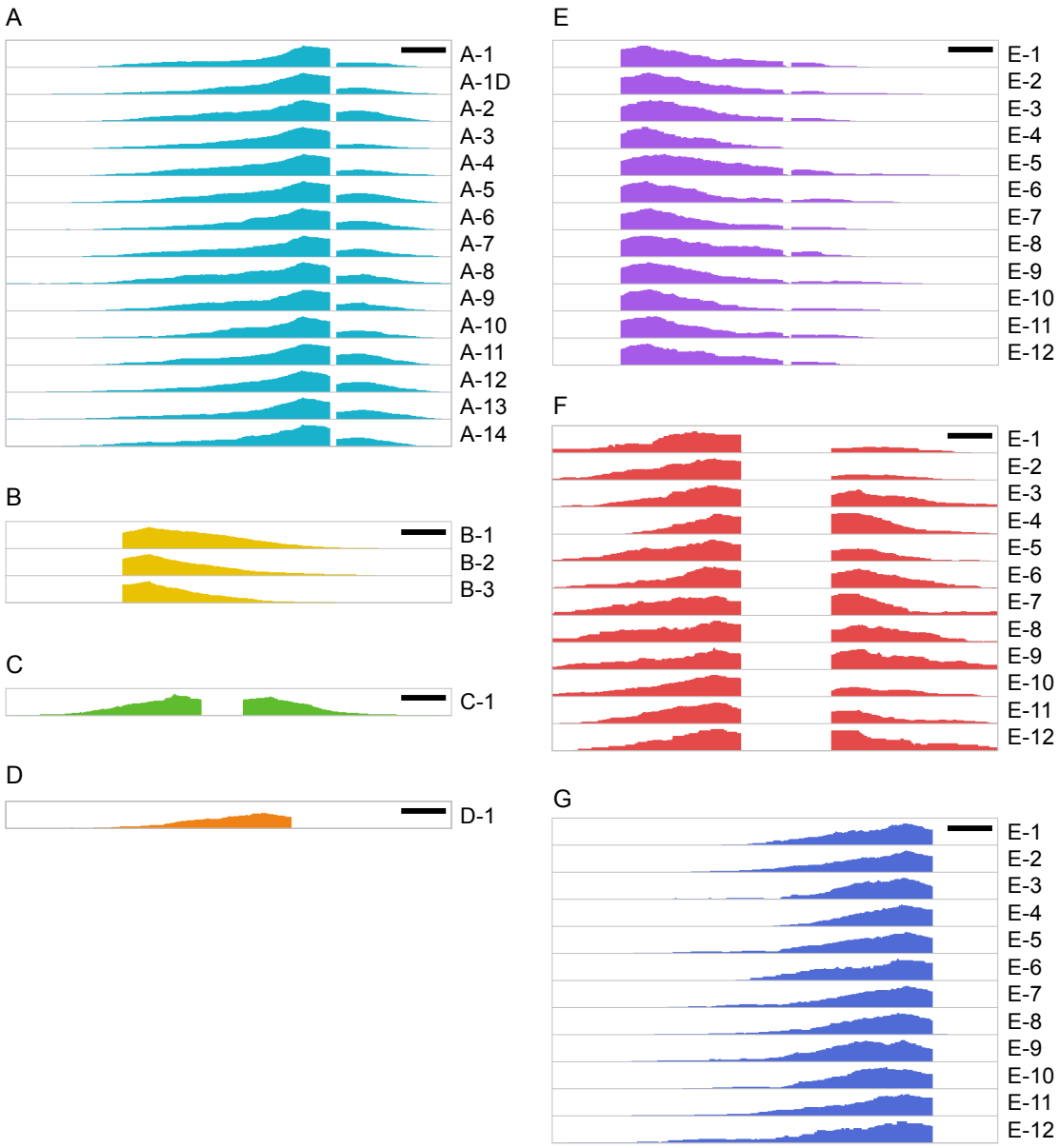

Figure 3-5. IGV read pileup of identified integration sites for each cell subline. Scale bar represents 50bp and read pileups are colored by scaffold. Loci are listed from left to right as they appear on the pileup. (A) Integration loci 1a & 1b on Scaffold T (Chromosome 1) for cells from lineage A. (B) Integration locus 2 on Scaffold U (Chromosome 2) for cells from lineage B. (C) Integration locus 3 on Scaffold V (Chromosome 6) for cell line from lineage C on the left and unidentified integration locus on the right. (D) Integration locus 4 on Scaffold W (Chromosome 5) for cell line from lineage D. (E) Integration locus 5 on Scaffold X (Chromosome 3) for cells from lineage E. (F) Integration loci 6a & 6b on Scaffold Y (Chromosome 4) for cells from lineage E. (G) Integration locus 7 on Scaffold Z (Chromosome 4) for cells from lineage E.

3.5 Conclusions

The method presented here allows for rapid verification of clonality or common lineage between different cell lines. Use of sequence capture increases the relevant information extracted from sequencing, and the bioinformatic analysis is rapid for processing the low number of reads required from each sample. Additionally, through the use of capture probes targeting common regions of the host vector, future clones can easily be added to the analysis without the need for new probe design. In general, a longer sequencing read length and an increased library size for each sample may have improved this dataset and would be helpful for detecting rarer integration events. The library size may be increased either through reducing the number of samples pooled before capture, or by increasing the input DNA from each sample to reduce PCR duplicates from library preparation. Despite some of these issues in sequencing, the pipeline was robust in its ability to distinguish cell lines of different lineages based on their integration sites. This method is thus a valuable, accessible tool to address clonality verification.

4 Recurring genomic structural variation leads to clonal instability and loss of productivity

Reproduced with permission from: Bandyopadhyay, A. A.*, O'Brien, S. A.*, Zhao, L., Fu, H. Y., Vishwanathan, N., & Hu, W. S. (2019). Recurring genomic structural variation leads to clonal instability and loss of productivity. *Biotechnol Bioeng, 116*(1), 41-53. doi:10.1002/bit.26823 *AAB and SAO contributed equally.

AAB analyzed CGH and whole genome sequencing data. SAO performed integration site analysis, chromosome analysis, qPCR, and qRT-PCR. AAB and SAO performed data analysis and visualization of the results.

4.1 Summary

Chinese Hamster Ovary (CHO) cells, commonly used in the production of therapeutic proteins, are aneuploid. Their chromosomes bear structural abnormality and undergo changes in structure and number during cell proliferation. Some production cell lines are unstable and lose their productivity over time in the manufacturing process and during the product's life cycle. To better understand the link between genomic structural changes and productivity stability, an Immunoglobulin G (IgG) producing cell line was successively single cell cloned to obtain subclones that retained or lost productivity, and their genomic features were compared. Although each subclone started with a single karyotype, the progeny quickly diversified to a population with a distribution of chromosome numbers that is not distinctive from the parent and among subclones. Comparative genomic hybridization (CGH) analysis showed that the extent of copy variation of gene-coding regions among different subclones stayed at levels of a few percent. Genome regions that were prone to loss of copies, including one with a product transgene integration site, were identified in CGH. The loss of the transgene copy was accompanied by loss of transgene transcript level. Sequence analysis of the host cell and parental producing cell showed prominent structural variations within the regions prone to loss of copies. Taken together, we demonstrated the transient nature of clonal homogeneity

in cell line development and the retention of a population distribution of chromosome numbers; we further demonstrated that structural variation in the transgene integration region caused cell line instability. Future cell line development may target the transgene into structurally stable regions.

4.2 Introduction

Chinese hamster ovary (CHO) cells are industrial workhorses for the production of recombinant protein therapeutics, such as monoclonal antibodies and Fc-fusion proteins, which require proper folding and post-translational modifications for their biological activity [60, 61]. The production cell line for these biologics is traditionally generated by random integration of the product transgene into the host CHO cell followed by transgene amplification and screening for high producing cell clones [44]. In addition to a high productivity, the cell line must also sustain its productivity, not only during the manufacturing process but also throughout the product's life cycle. To mitigate risks related to genetic changes in the producing cell line, single cell cloning is performed prior to the establishment of the cell stock to ensure the homogeneity of the starting cell population [45, 62]. This minimizes the probability that a subpopulation of cells overtakes the population, possibly causing changes in the productivity or product quality.

For normal diploid cells, such as many different types of stem cells, single cell cloning ensures the homogeneity of the ensuing cell population. However, aneuploid cell lines, including CHO cells, have abnormal chromosome number and structure. During proliferation, they continuously undergo genomic changes such as mutations, deletions, duplications, and other structural alterations due to errors in DNA replication and repair, and mistakes in chromosome segregation. As a result, these cells have a wide distribution of chromosome number, which has been shown in commonly used cell lines such as HEK293 [63], MDCK [64, 65], and Vero cells [24, 66, 67]. This heterogeneity in chromosome number and structure has also been demonstrated in CHO cells [23, 68-71].

For a production CHO cell line, a large number of cell divisions are required to expand the cell population and have enough cells to fill a manufacturing bioreactor and to create enough cell banks to encompass a product's life cycle. The subsequent accumulation of genome aberrations over time can lead to genetic and phenotypic heterogeneity among

CHO cells, even those which are clonally derived [72]. This heterogeneity can occur in the form of genomic and epigenomic variation [73, 74] or changes to cell phenotype or productivity [75, 76].

There are a number of reported mechanisms leading to production instability [77]. These include loss of transgene copy number [78, 79], promoter methylation [80, 81], and other epigenetic silencing mechanisms in the promoter region [82]. In most studies, gene amplification was used to acquire multiple copies of transgene. How the loss of transgene copy number is affected by structural changes in the surrounding genome regions has not been elucidated.

Using Comparative Genome Hybridization (CGH), it has become possible to globally survey copy number changes in genomic loci within CHO cells [83]. High throughput sequencing technologies have also greatly facilitated whole genome sequencing and the detection of structural changes in the genome. The combination of these tools with classical karyotyping and chromosome counting allows us to evaluate the effect of genome stability on phenotype stability in CHO cells. We chose to study productivity because it is the most important phenotype of an industrial cell line, and because it is readily measurable after single cell cloning.

We first eliminated the population heterogeneity of a recombinant DG44 cell line by single cell cloning to obtain subclones of high and low productivities, then allowed the heterogeneity to reestablish. The genomic heterogeneity of the populations derived from high and low producing subclones was investigated at the macroscopic level using karyotyping and chromosome counts, and at the microscopic gene and sequence levels using genome sequencing and CGH data. We also identified the primary transgene integration sites in the genome and examined local genomic alterations at the integration loci.

4.3 Materials and methods

4.3.1 Cell lines and culture conditions

A recombinant CHO-DG44 cell line producing IgG (rDG_IgG) was used in this study. rDG_IgG and all derived clones were grown in T-flasks (Corning) in CHO-S-SFM-

II Medium (Gibco). Cells were incubated at 37°C in a humidified incubator with 5% CO_2. Single cell cloning was performed by limiting dilution at a seeding density of 0.2 cells per well.

4.3.2 Preparation of metaphase cells and Chromosome counting

Approximately 5 x 10^6 cells in exponential phase of growth were submitted to the Cytogenomics core at the University of Minnesota for metaphase spreads and G-banding analysis. Briefly, cells were cell cycle arrested by colcemid treatment, followed by incubation in 0.075M KCl at 37°C. Cells were fixed in a 3:1 solution of methanol:acetic acid, and then dropped onto slides. Slides were treated with a trypsin solution for improved banding. Wright's stain was used for G-banding and visualization of chromosomes. Imaging of metaphase spreads was done on a Nikon Diaphot using a 100x oil objective. At least 70 chromosome spreads were counted per sample. Chromosome counting was done manually using ImageJ (NIH).

4.3.3 Genome sequencing

Genomic DNA from H_1 was extracted using DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA) and paired-end sequenced (101 bp) using Illumina Hi-seq 2000. Sequencing data for the DG44 genome was obtained from Lewis, Liu [84] (Accession number: SRS406582). Raw sequencing reads were quality-trimmed and were mapped to an annotated genome, UMN2.0 [85] using the gap-enabled aligner BWA mem [86] in paired-end mode. The SAMtools suite of algorithms (v.0.1.18) [87] was used to process the resulting alignments. Reads with a minimum paired-end mapping quality of 20 were used. Structural variant analysis was performed using DELLY [88].

4.3.4 Comparative genomic hybridization (CGH)

The construction of a microarray for comparative genomic hybridization (CGH) based on the assembled ESTs for Copy Number Variation (CNV) analysis has been described previously. [83] Since the original manufacturer of the array no long provide the oligoDNA array, a new 4-plex array each with 167,508 DNA probes was custom designed and manufactured by Agilent. The array focused on the transcript coding regions of the genome.

For CGH array hybridization, genomic DNA was extracted from either CHO cell lines or Chinese Hamster liver tissue using DNeasy Blood & Tissue Kit (Qiagen, Vaencia, CA) or NucleoSpin® Tissue Kit (Macherey-Nagel, Duren, Germany). DNA was quantified using a NanoDrop 1000 Spectrophotometer and the genomic DNA integrity was assessed in a 1% agarose gel. The DNA pre-processing, labeling and hybridization to the Agilent microarray was done by the genomics core facility at the University of Iowa. Briefly, the DNAs were labeled using the SureTag DNA Labeling Kit followed by hybridization to a 4x180K 2-dye expression microarray from Agilent (Design ID 075155) as per manufacturer's instructions. The arrays were scanned using Agilent SureScan Microarray Scanner G2600D. The raw intensity data from the microarray hybridization images were obtained from Agilent CytoGenomics 3.0 software in form of a .txt file and was normalized by LOWESS method. Subsequent analysis was performed using MATLAB® (version 2015b; MathWorks). In order to identify segments of the genome with variation in copy number relative to liver, a circular binary segmentation-based DNAcopy algorithm was used to identify segments affected by copy number variation [89, 90]. This algorithm uses statistical analysis to smooth out noisy CGH probe data and call discreet regions of the genome as having a particular copy number.

4.3.5 Integration site analysis

Sequence capture method

Integration site analysis for rDG_IgG was performed using sequence capture. A diagram of this process is shown in Figure 4-1. Genomic DNA (gDNA) was sheared, and a size selection was performed to isolate DNA in the 300-500bp range. An indexed and universal adapter were ligated to either end of the sheared fragments. The Agilent SureSelect Enrichment System was then used to enrich for fragments containing vector sequence. Short biotinylated RNA probes (120bp) were designed to tile along the vector sequence with complete coverage. These probes were hybridized to the gDNA, and streptavidin-coated beads were used to pull down fragments containing vector sequence. After washing the beads and digesting the RNA probes, the captured fragments were purified, and sequenced. Sequencing was performed using single end Illumina MiSeq, with an average read length of 250bp. Reads from sequencing were first mapped to the vector

to filter out reads without vector information, and then mapped to the genome to filter out vector only reads. Split reads and pileup information was used to determine the integration junction.

PCR-based method

Integration site analysis for rDG_IgG was also performed using a nested PCR method. A diagram of this process is shown in Figure 4-2. gDNA was extracted from H_1 cells using NucleoSpin Tissue Kit (Takara Clontech). The gDNA was then digested with three restriction enzymes: DraI, SspI, and HpaI. Digested DNA was further purified using NucleoSpin Gel and PCR Clean-Up Kit (Takara Clontech). After purification, GenomeWalker adapters (taken from the Lenti-X Integration Site Analysis Kit, Takara Clontech) were ligated to the fragmented gDNA. Two successive rounds of PCR were then performed to amplify gDNA fragments containing vector sequence using the Advantage 2 PCR Kit (Takara Clontech). In order to capture the different integration sites, multiple vector specific primers were used that spanned the product integration vector. First, PCR was done using one of primary vector specific primers and Adapter Primer 1 (AP1), which binds to the ligated adapter. The PCR reaction was diluted, and then added to a second reaction using a corresponding nested vector primer and Adapter Primer 2 (AP2), found downstream of the original set used for amplification. Gel electrophoresis was used to visualize the nested PCR reaction, and resulting bands were extracted from the gel and purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research). The purified PCR products were Sanger sequenced at the University of Minnesota Genomics Center (Saint Paul, MN).

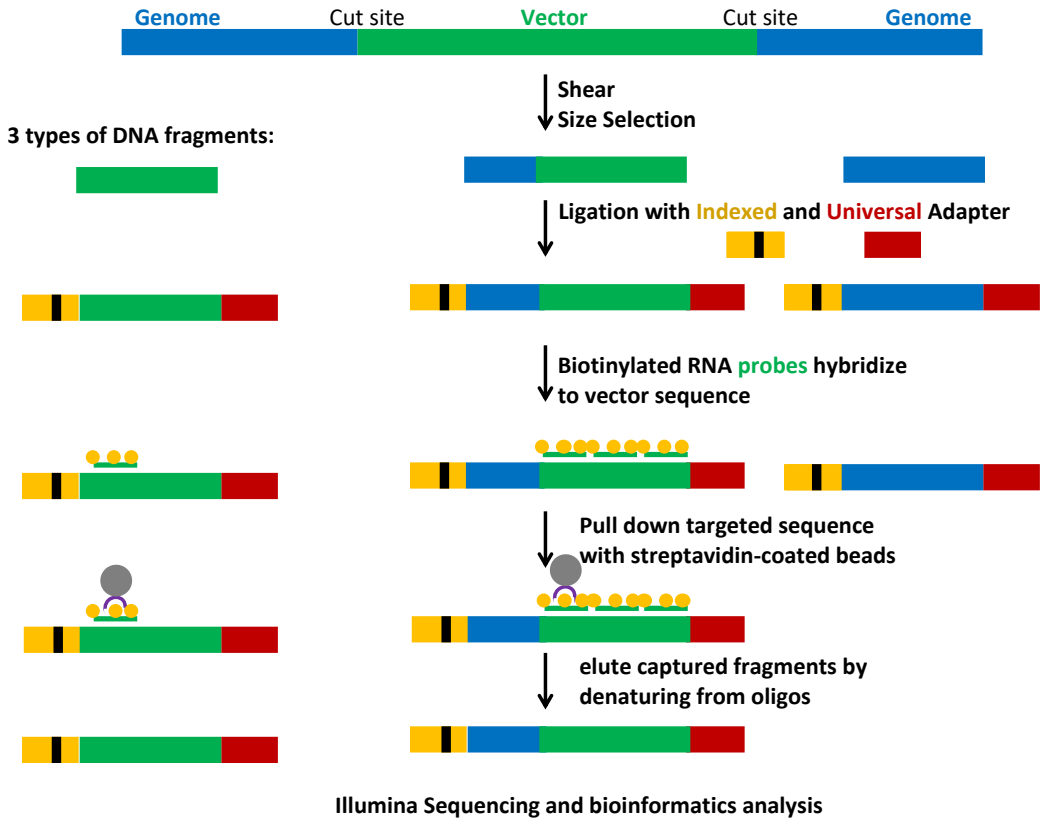

Figure 4-1. Sequence capture-based integration site analysis method. Genomic DNA was first sheared, and size selection was used to isolate DNA in the 300-500bp range. Illumina sequencing adapters were then ligated, and biotinylated RNA probes and streptavidin beads were used to capture fragments containing vector sequence. Isolated fragments were sequenced using Illumina sequencing. See materials and methods for a detailed description.

4.3.6 Genomic DNA quantitative PCR (qPCR)

Genomic DNA was isolated using NucleoSpin® Tissue Kit (Macherey-Nagel, Duren, Germany). The qPCR assay was conducted in triplicate using 20ng of genomic DNA in a 20 µL reaction volume on a CFX Connect Real-Time PCR Detection System (Bio-Rad, Hercules, CA) machine using SYBR Select Master Mix (Applied Biosystems, Foster City, CA). The cycle numbers obtained from the qPCR experiment were normalized to an unamplified region in the genome, β-actin, to account for differences in DNA loading. Data is reported as fold change relative to parent. The raw cycle number for β-actin was similar for all cell lines.

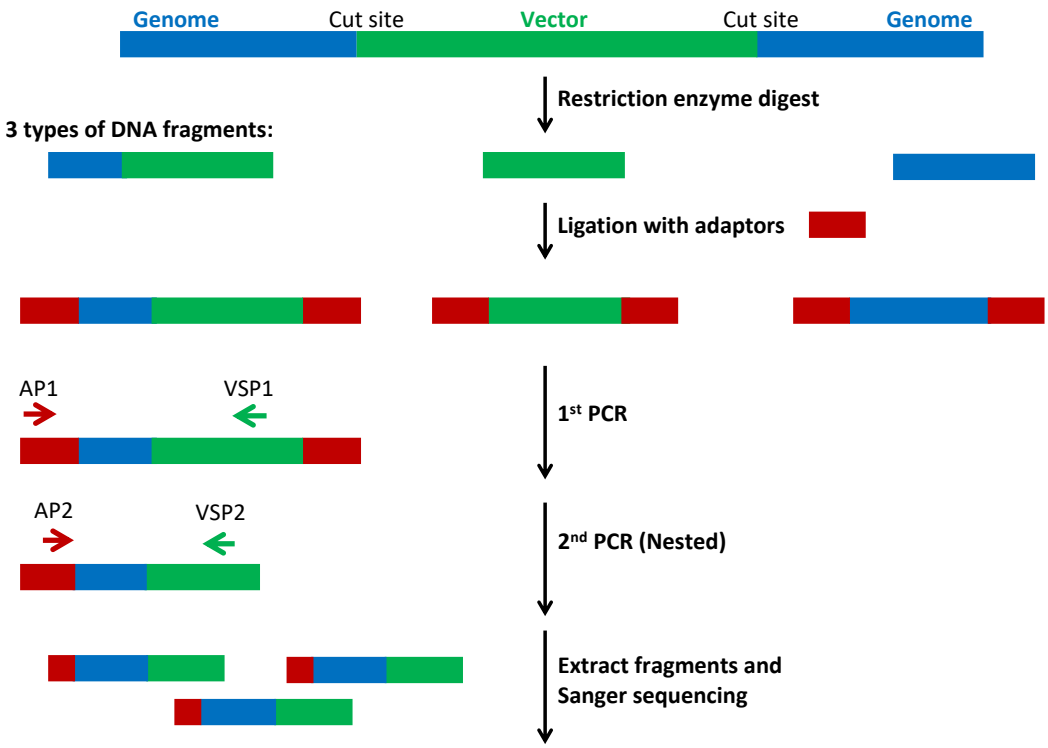

Figure 4-2. Nested PCR based integration site analysis method. Genomic DNA was first fragmented using restriction enzymes and then ligated with an adaptor. Two rounds of PCR using a vector specific primer (VSP) and an adaptor specific primer (AP) are used to enrich for fragments containing both vector and genome sequences. Extracted fragments are then sequenced using Sanger sequencing. See materials and methods for a detailed description.

4.3.7 qRT-PCR for quantifying transcript levels

RNA was extracted from cells using RNeasy Mini Kit (Qiagen), and cDNA synthesis was done using SuperScript III First-Strand cDNA Synthesis SuperMix. qRT-PCR was performed using SYBR Select Master Mix (Applied Biosystems) on a CFX Connect Real-Time PCR Detection System (Bio-Rad, Hercules, CA) using manufacturer recommended standard protocol. qRT-PCR data for the subclones was normalized to GAPDH, and then to their respective parents. Data is reported as fold change relative to parent.

4.4 Results

4.4.1 Derivation of subclones

A CHO cell line (rDG_IgG) that has been extensively cultured in the laboratory [91-93] was subcloned and IgG productivity was quantified (Figure 4-3). The highest producing clone, H_1, was isolated and expanded for a second round of subcloning 35 days

after the first subcloning (~ 30 population doublings). The two subclones with highest and lowest IgG titer derived from H_1 (designated as H_1H_2 and H_1L_2) were again expanded to obtain enough cells for further characterization. H_1H_2 was then subcloned for the third round 55 days after the second subcloning (~40 population doublings). The highest and lowest producing clones were again obtained and designated as $H_1H_2H_3$ and $H_1H_2L_3$ respectively (Figure 4-3). The growth rates of the derived clones were all similar.

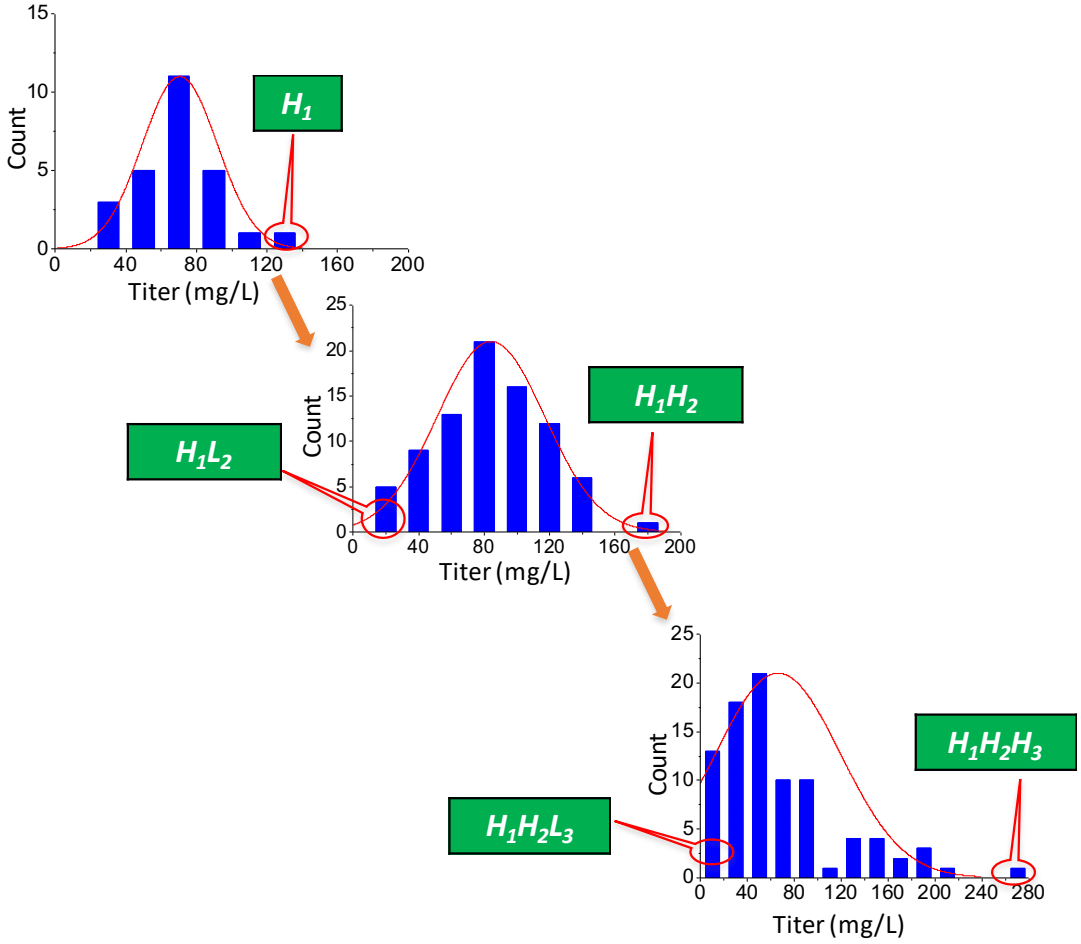

Figure 4-3. Derivation of cell lines through subcloning and the titer distribution.

4.4.2 Rapid rise of heterogeneity in karyotype and chromosome number

The inbred female Chinese hamster from which CHO cells were derived had 11 pairs of chromosomes (classically the chromosomes were numbered 1, 2, X, 4 until 11, skipping chromosome 3) [69]. Figure 4-4A shows representative cell karyotypes from H_1H_2 and $H_1H_2H_3$ based on the classical assignment of Chinese hamster chromosomes. Each

metaphase spread exhibited a few normal appearing chromosomes as well as some abnormal ones. The copy number of each normal chromosome in the H_1H_2 and $H_1H_2H_3$ cells analyzed is shown in Figure 4-4B. Each line represents the chromosome count for one metaphase spread analyzed. More than half of the cells retained a very similar distribution of normal chromosomes for both H_1H_2 and $H_1H_2H_3$. Chromosomes 6 retained two copies in all cells. No cell was found to lose all copies of chromosome 1, while some trisomic cells were seen.

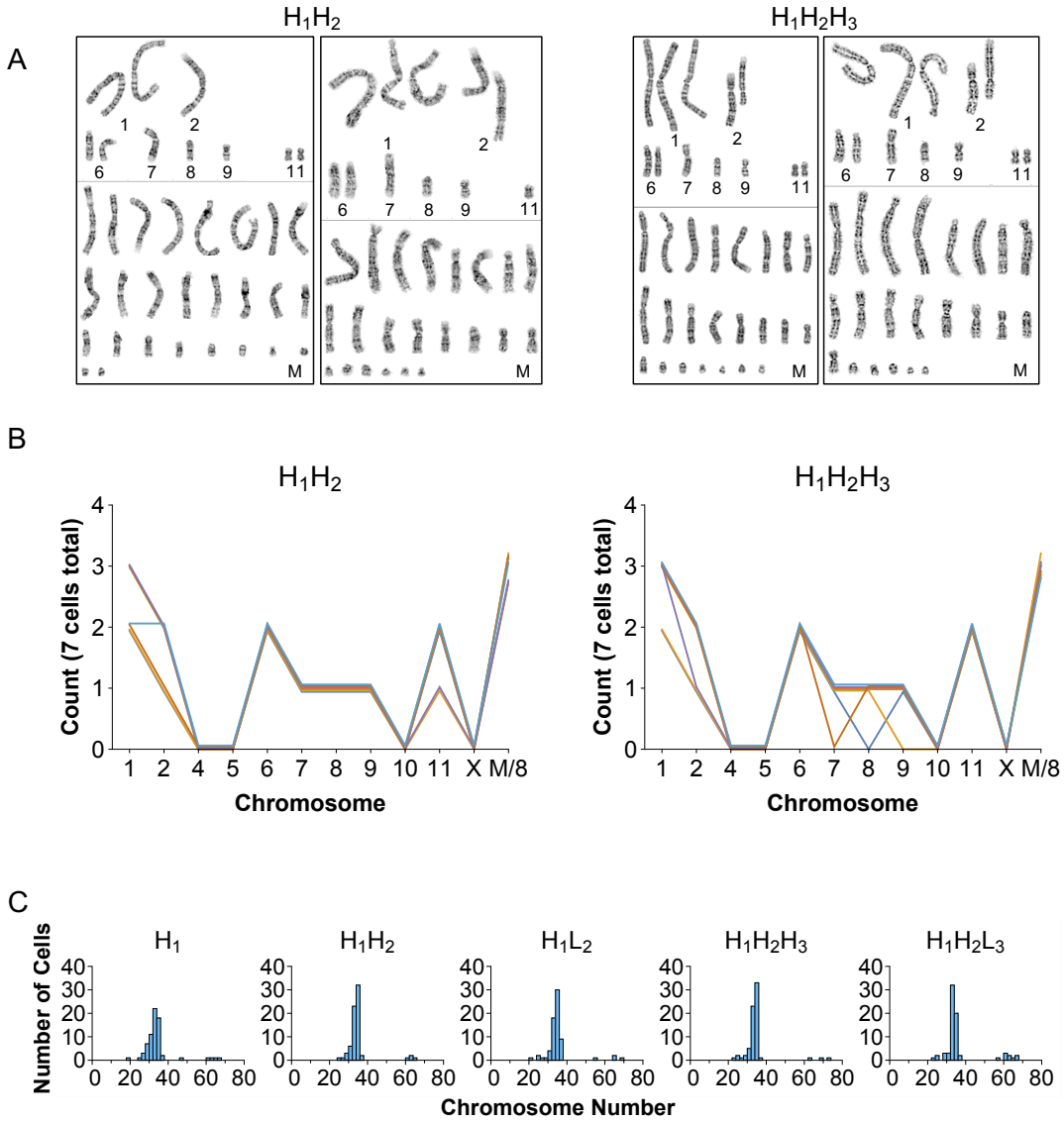

Figure 4-4. Karyotypes and chromosome counts of subclones. (A) Representative karyotype images from H_1H_2 and $H_1H_2H_3$. (B) Distribution of all the chromosomes from the karyotypes. The marker chromosome numbers are divided by 8 for visualization purposes (M/8). (C) Chromosome number distributions for all the subclones obtained from ~70 metaphase spreads.

To survey a large number of cells, we utilized chromosome counting of metaphase spreads instead of karyotyping each chromosome. Chromosome numbers of approximately 70 cells were obtained for all five subclones. Representative images are shown in Figure 4-5. All the cell lines showed a pseudo-triploid chromosome set of ~35 chromosomes with broad cell-to-cell variance in chromosome number within the population (Figure 4-4C). A small number of cells in each cell line had a lower chromosome number (~15-25) while some had almost 70 chromosomes. The mode of chromosome number as well as the chromosome number distribution was largely similar among the five subclones. $H_1H_2L_3$ may have a slightly higher number of cells with a very high number of chromosomes compared to its sister $H_1H_2H_3$ and parent H_1H_2. Nevertheless, the data show that in the population doublings required to attain sufficient cells for systematic investigation, the once "clonal" population displayed a spread of chromosome numbers. Although the chromosome number and karyotype of each initial cell from subcloning was not known and were likely to be different from each other, the resulting populations had remarkably similar chromosome number distributions. The small sample size did not reveal any discernible difference in chromosome number in the phenotypically "abnormal" low-producing subclones.

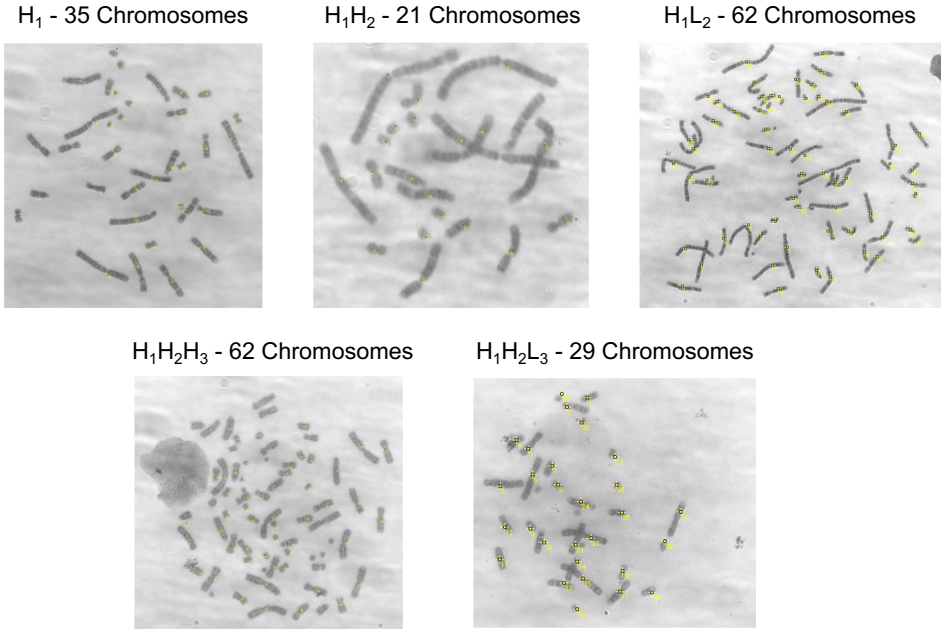

Figure 4-5. Representative images from metaphase spreads. Chromosomes counted from image are marked in yellow.

4.4.3 Gain and loss of gene copy number among subclones

To assess the genomic changes at a microscopic scale, a comparative genomic hybridization (CGH) array (focused on transcript-coding regions) was used to detect gain or loss of genomic regions. Since the reference DNA was from a diploid genome, the normalized signal ratio of the sample to the reference for a probe reflects the gain or loss of abundance level of the corresponding gene. It should be noted that the genomic structure of the cell population was heterogeneous, and in a given genomic region, some cells may have no gain or loss of copy, while a subpopulation might have gained or lost copies. It is also possible that cells with multiple structural changes in the same gene region may appear to be distinct copy number change events, or that different degrees of copy number change might occur in a region that is prone to structural changes. Thus, the gene copy number data may not show discrete changes in copy numbers (e.g. one allele changed or both alleles changed) as in the case of a homogeneous population with a diploid genome.

Pairwise comparisons of CGH data for the subclones are shown in Figure 4-6. The dotted horizontal and vertical lines a, b mark one copy gain with respective to liver ($\log_2 1.5 = 0.58$ i.e. 3 copies instead of 2), while the lines c, d mark one copy loss with respect to liver ($\log_2 0.5 = -1$ i.e. 1 copy instead of 2 copies). 5,072 (3.02 %) probes had gained copy(-ies) in both H_1 and H_1H_2 as compared to liver (upper yellow box, Figure 4-6A). Similarly, 932 (0.55 %) probes lost copy(-ies) for both H_1 and H_1H_2 (lower yellow box, Figure 4-6A). Similar numbers of about 5000 for gain of copy(-ies) and 900 for loss of copy(-ies) were seen in the other comparisons. Despite the seemingly large change of chromosome number from normal diploidy, the genome showed a relatively small gain and loss of copies at the gene probe level compared to a diploid reference.

The 45° solid red lines (x=y) mark the probes for which the two samples have the same copy number relative to liver. The dotted red lines e, f indicate the bounds for 1.5-fold gain or loss from one cell line to the other. The vast majority of probes lie within the bounds of 1.5-fold change (region bounded by red dashed lines). 959 (0.57 %) probes (outside the upper red dashed line labeled e) showed a higher copy number in H_1H_2 compared to its parent H_1 (Figure 4-6A), whereas 280 (0.17 %) probes showed a lower copy number in H_1 than H_1H_2 (outside the lower red dashed line labeled f). Similarly, 1199

probes showed a gain in H_1L_2 compared to its parent H_1 and 753 probes showed a loss in H_1L_2 compared to H_1 (Figure 4-6B). $H_1H_2H_3$ compared to H_1H_2 showed a very narrow spread of probe intensities (82 probes are higher and 35 probes are lower, Figure 4-6D), indicating that these two samples are very similar at the gene copy abundance level, whereas $H_1H_2L_3$ and its parent H_1H_2 showed a higher spread, with 271 probes higher and 786 probes lower in $H_1H_2L_3$, (Figure 4-6E).

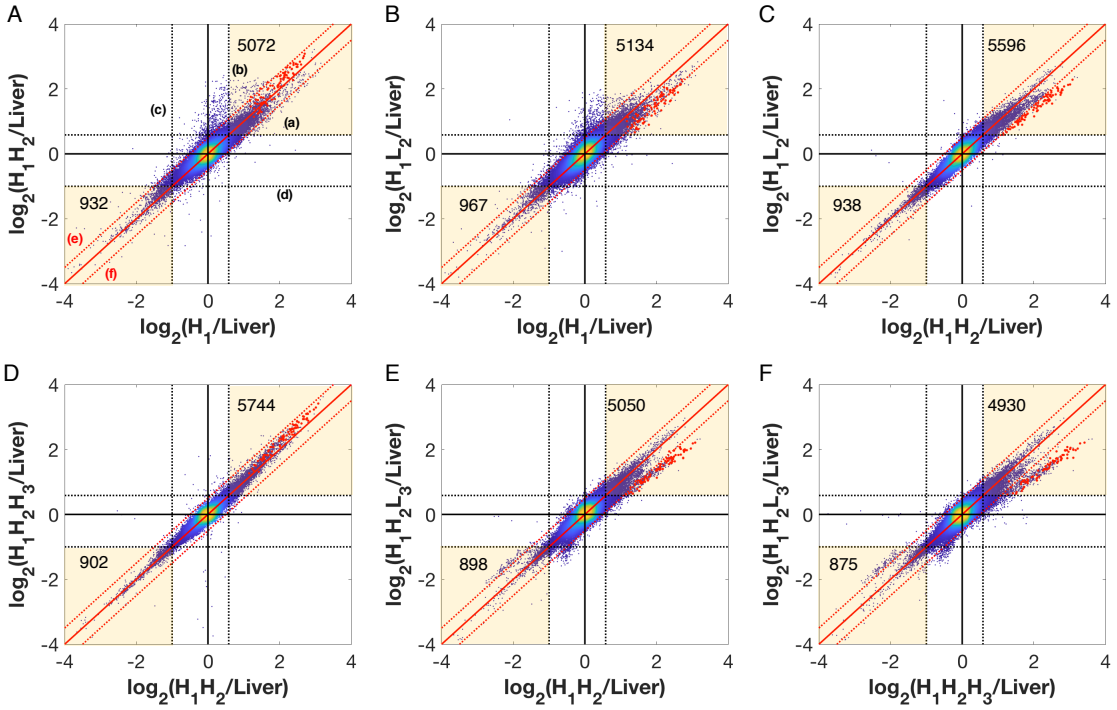

Figure 4-6. Pairwise comparison of $\log_2$(Sample/Liver) probe intensities for the 5 subclones. The dotted black lines (labeled a and b), both horizontal and vertical, are the bounds for one copy gain with respect to liver ($\log_2 1.5 = 0.58$) (i.e. 3 copies instead of 2). The dotted lines (labeled c and d) mark the bounds for loss of one copy with respect to liver ($\log_2 0.5 = -1$), i.e. 1 copy instead of 2 copies. The solid red line along the diagonal indicates the line x=y which implies both samples have the same copy number. The red dashed diagonal lines (labeled e and f) indicate the bounds for 1.5-fold difference in copy between the two cell lines. The red dots are the probes belonging to a genomic 4 Mbp region which show loss during high-to-low producer transition.

Figure 4-6C shows the comparison of two sublines derived from the same parent, H_1H_2 and H_1L_2, where 145 (0.09 %) probes showed an increased copy number in H_1H_2 over H_1L_2 while 87 (0.05 %) probes showed a lower copy number. $H_1H_2L_3$ showed higher divergence from its sister clone $H_1H_2H_3$ (276 probes ↑ and 1281 probes ↓, Figure 4-6F). Interestingly, both low producers (H_1L_2 and $H_1H_2L_3$) had ~100 probes with gained copy

from the diploid reference show a decrease in copy number from their respective high producing sister clones (H_1H_2 and $H_1H_2H_3$) (Figure 4-6C and F).

4.4.4 CNV between high and low producing clones

Among the probes which showed consistent loss during high-to-low producer transitions, ~85 probes (Figure 4-6, red dots) belonged to a 4 Mbp region within the genome, flanked by the genes *Csmd3* and *Rspo2* (Figure 4-7A, yellow box). Each point represents the ($\log_2$) intensity ratio of the sample to diploid genome (liver) of a probe lined up according to its position in the genome region. The black line denotes the mean ($\log_2$) intensity ratio of a contiguous segment computed by DNAcopy. The region marked in yellow is amplified in the parent H_1. The high producing clones, H_1H_2 and $H_1H_2H_3$, retained the gain of copy number seen in the parent, but the low producers H_1L_2 and $H_1H_2L_3$ showed a loss in copy number from their respective parents, H_1 and H_1H_2. Importantly, the loss of the amplified segment happened in two independent transition events from a high producing parent to low producing clones.

We further investigated this region using whole genome sequencing read pileups from H_1 and the host cell line DG44 (Figure 4-7B). The pileup of sequencing reads in this region for the host line DG44 showed almost a uniform sequence depth in the entire genomic region except a few highly amplified spikes at the left end. In contrast, the pileup for H_1 showed increased depth in read pileups over the entire region. Notably, a nearly 4 Mbp region was further amplified to different levels in three segments. This is consistent with the DNAcopy call of ~4 Mbp segmental gain of copy from the CGH data. The CGH data lacked the spatial resolution to divide the segment further into different levels of amplification as the CGH microarray probes covered only the gene coding regions. Nevertheless, the sequence read pileup confirmed the CGH finding that the repeatedly lost 4 Mbp region was amplified in H_1.

We investigated the stability of this region using the CGH data reported previously for other cell lines [83]. In CHO-K1, almost the entire segment has a loss of ~1 copy from diploid ($\log_2 0.5 = -1$ i.e. 1 copy instead of 2 copies, Figure 4-8A). The loss was retained by K1 derived producing cell lines rK1_IgG and rK1_2C10 (Figure 4-8B and C) and was also seen in DXB11 (Figure 4-8D). However, the $\log_2$ segment mean relative to diploid of

the DXB11 derived cell line rDX_Fcf was between 0 and -1 (Figure 4-8E). This might indicate a heterogeneous population in which some cells retained the loss while some others gained a copy relative to DXB11. Interestingly, the other DXB11 derived cell line, rDX_IgG_32, gained a copy relative to DXB11 and returned to the level of diploid cells (Figure 4-8F). DG44 shows a normal ploidy state (Figure 4-8G), consistent with the sequencing pileup data. In the DG44 derived lines, different segments of this region were amplified to varying extents as illustrate by the plots of H_1L_2 and $H_1H_2L_3$ (Figure 4-7A).

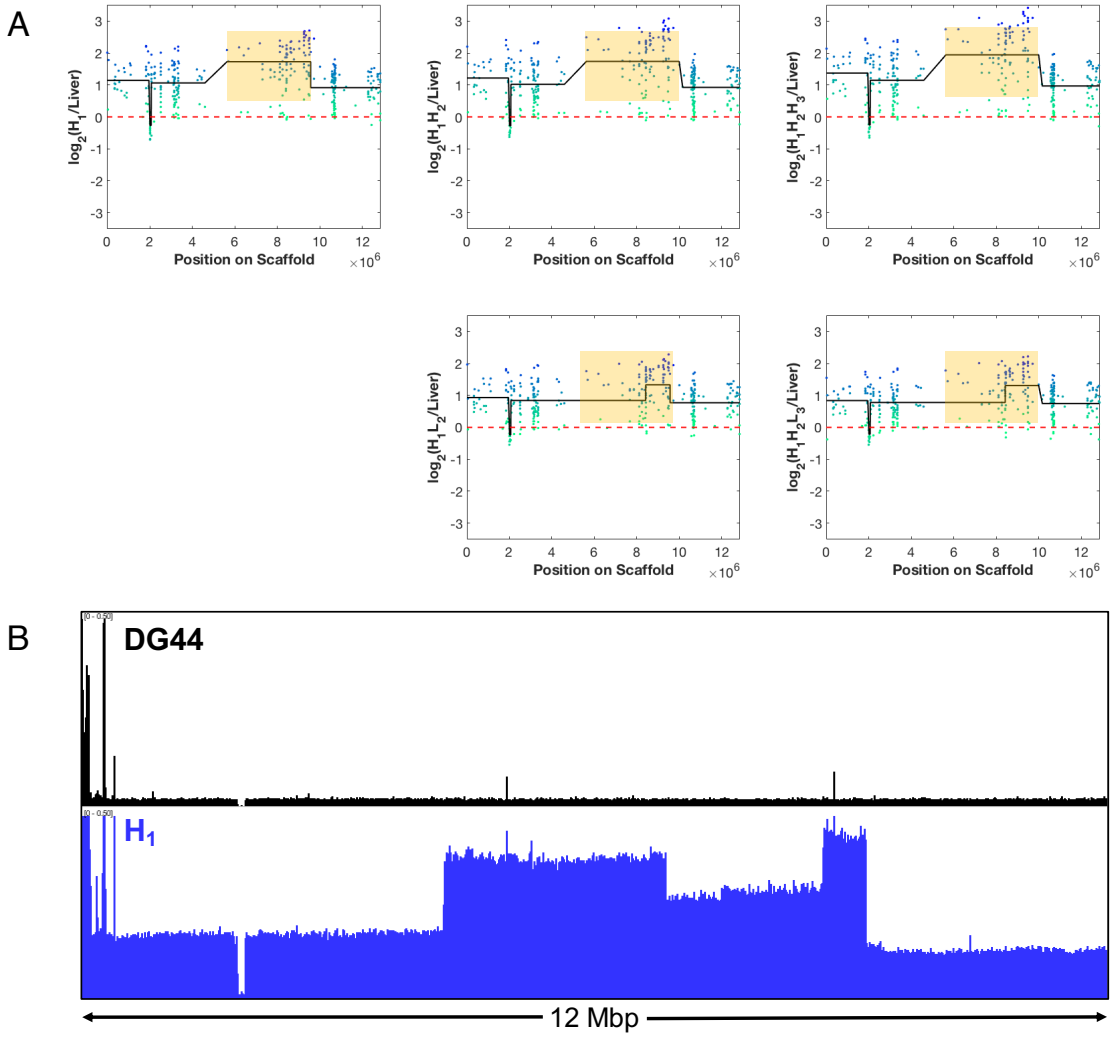

Figure 4-7. Loss of an amplified genomic region in low producers. (A) Two independent losses of a 4 Mbp genomic region seen during high-to-low producer transition. The green-blue dots represent the log-ratio intensities for each probe in the region relative to liver. The black line is the mean intensity of the segments as identified by the DNAcopy. (B) Normalized sequencing read pileup (Reads per Million) for DG44 and H_1. H_1 shows amplification of the entire genomic region, correlating with the CGH data.

Overall, the data suggests that this region of genome is not only prone to change in copy number over a long stretch of segments as seen in the transition of high producing cells to low producing cells but has also been deleted or amplified in various CHO cell lines. This further suggests that some genome regions may be more prone to structural changes.

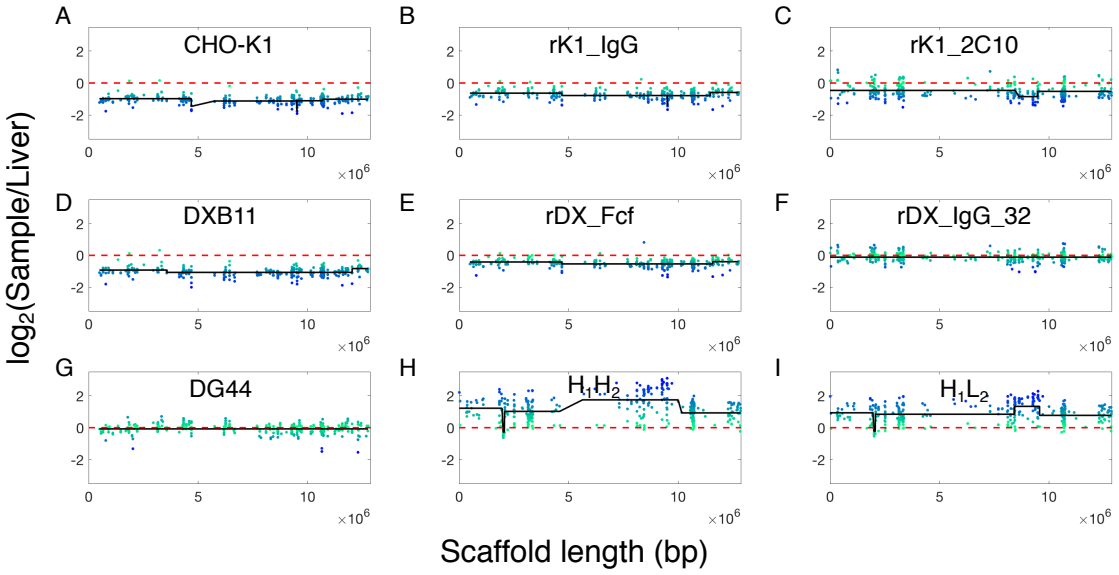

Figure 4-8. Copy number variation in the genomic scaffold (shown in Figure 4-7) in cell lines from other lineages. (A) CHO-K1 and derived cell lines rK1_IgG (B) and rK1_2C10 (C). (D) DXB11 originally derived from CHO-K1 and its derived cell lines rDX_Fcf (E) and rDX_IgG_32 (F). (G) DG44 and its derived cell lines H_1H_2, H_1L_2. The green-blue dots represent the log-ratio intensities for each probe in the region relative to liver. The black line is the mean intensity of the segments as identified by the DNAcopy.

4.4.5 Identification of transgene integration sites

Having shown that the occurrence of low producing clones was accompanied by recurring loss of copies in some genome regions, we identified the transgene integration sites to determine possible gain or loss of copies in those regions. Three methods were employed to identify the transgene integration sites in the genome. First, biotinylated probes were used to capture the genome fragments containing sequences of different regions of the vector (Figure 4-1). The isolated fragments were then sequenced using Illumina sequencing. Due to the short-read length of Illumina MiSeq (250bp), the confidence for mapping of split reads was low in some cases. From reads containing both vector and genome sequences, three integration sites were identified (Table 4-1). The read

pileup showed a blunt boundary at the vector-genome junction and trailing read depth moving away from the integration site as expected (Figure 4-9). Next, vector specific primers designed along the vector sequence were used to amplify DNA segments extending beyond the genome vector junction (Figure 4-2). Amplified fragments were then separated on a gel (Figure 4-10), extracted, and sequenced using Sanger sequencing. The longer genome sequences obtained gave higher confidence in the integration site identified. Using this method, we confirmed the previously determined integration sites and identified an additional site (Table 4-1). Further, DELLY was used on whole genome sequencing data of H_1 to identify split reads that map to both the genome and the vector. Based on the paired-end and split reads spanning the integration junction, all but one of the integration sites found using the PCR-based method were identified, and a fifth site was also identified. The missing integration site within the intron of the *Rc3h1* gene, as discussed later, had complex sequence rearrangements that made it difficult to identify. Combining the three methods we were able to identify the five transgene integration sites with high confidence (Table 4-1).

Table 4-1. Description of integration sites in rDG_IgG and methods used for identification. Prime (‘) denotes other side of integration site.

Site #	Integration Location	Vector Junction	Region Amplified	Detection Method		
				PCR + Sanger	DELLY	Sequence Capture
1	Rc3h1, intron	CMV	Yes	✓	✗	✓
2	Vps13b, intron	**β**-Lac	Yes	✓	✓	✓
2’	Vps13b, intron	SV40 pA	Yes	✓	✓	✗
3	IGR	KanR	No	✗	✓	✓
3’	IGR	KanR/ SV40 pA	No	✗	✓	✗
4	Stag2, intron	**β**-Lac	No	✓	✓	✗
5	CYP3A31, intron	DHFR	No	✗	✓	✗
5’	CYP3A31, intron	DHFR	No	✗	✓	✗

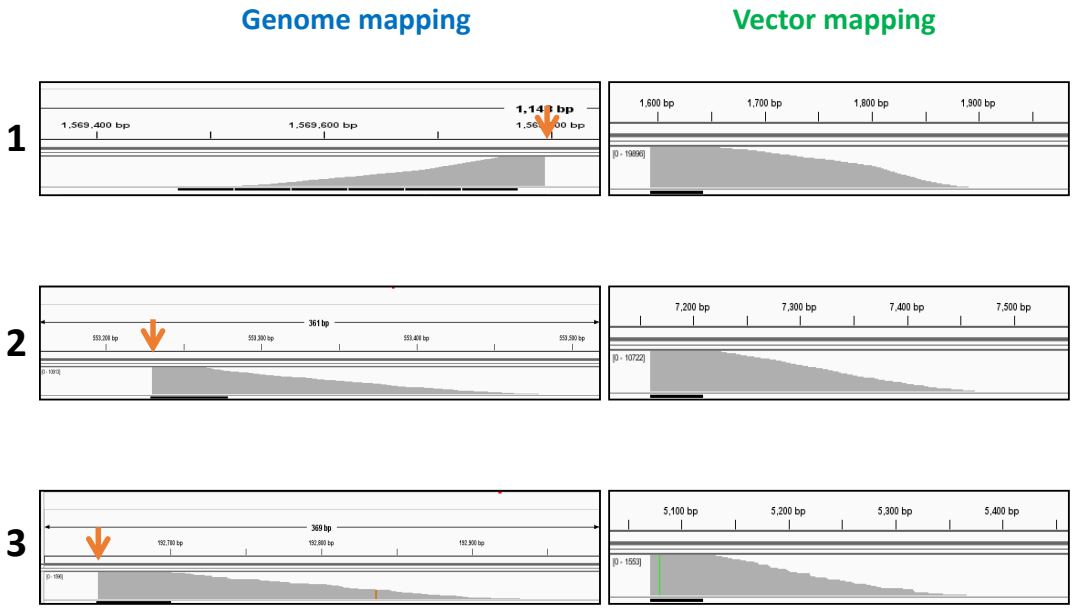

Figure 4-9. Read pile up for sequence capture based integration site analysis. Orange arrows denote integration junctions.

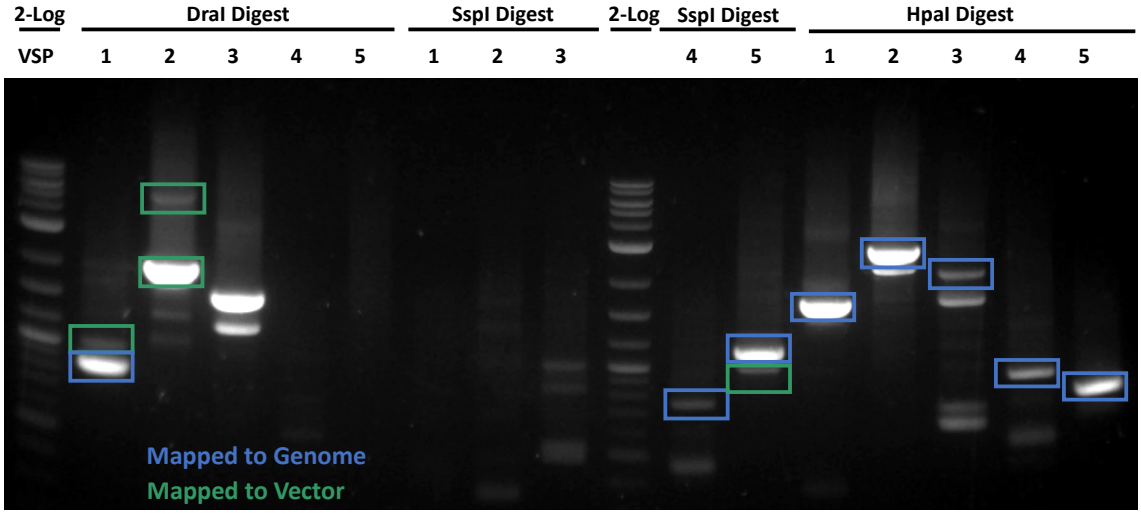

Figure 4-10. Gel electrophoresis of nested PCR for PCR-based integration site analysis. Extracted fragments are denoted with a box. (Blue: mapped to genome; Green: mapped to vector). Restriction enzyme library used to generate fragments is listed above lanes. VSP1-5 denote the vector specific primer used to amplify the fragments. (2-Log: 2 Log Ladder, NEB).

4.4.6 Loss of transgene in high-to-low producer transition

We next examined the CGH data on the integration sites identified for changes in copy number. The relative probe intensity at or near all integration loci was similar among all subclones except for the probe representing *Rc3h1* (site #1 in Table 4-1). *Rc3h1* had a much higher copy number than the diploid genome, suggesting that the locus had been amplified (Figure 4-11A, marked in red circle). The copy number appeared similar in H_1 and H_1H_2, but slightly higher in $H_1H_2H_3$. However, the copy number was markedly reduced in low producing clones. Note that H_1H_2 and H_1L_2 were both derived from H_1, while $H_1H_2H_3$ and $H_1H_2L_3$ were both derived from H_1H_2. The data thus showed that the *Rc3h1* locus lost copy in two independent events associated with the emergence of low producers. To examine whether the copy loss of *Rc3h1* was accompanied by the loss in transgene copy, the copy numbers of the IgG heavy chain and light chain genes in the clones were quantified using qPCR (Figure 4-11B). The high producing daughter cell line H_1H_2 retained the transgene copy number from its parent H_1 whereas H_1L_2 showed a loss of IgG heavy chain copy compared to its parent H_1. High producing daughter cell line $H_1H_2H_3$ showed an increased copy number of IgG heavy chain compared to its parent H_1H_2, consistent with the increased probe intensity for *Rc3h1*, the integration site. The low producing daughter cell line $H_1H_2L_3$ showed a loss in copy number of IgG heavy chain compared to its parent H_1H_2. IgG transcript expression from qRT-PCR also showed a similar trend where the low producing clones showed a 6 to 15-fold decrease in the expression of IgG heavy chain (Figure 4-11C). These results imply that the loss of productivity is likely to be a consequence of the loss of an IgG heavy chain gene copy at the *Rc3h1* locus and subsequent loss in transcript expression.

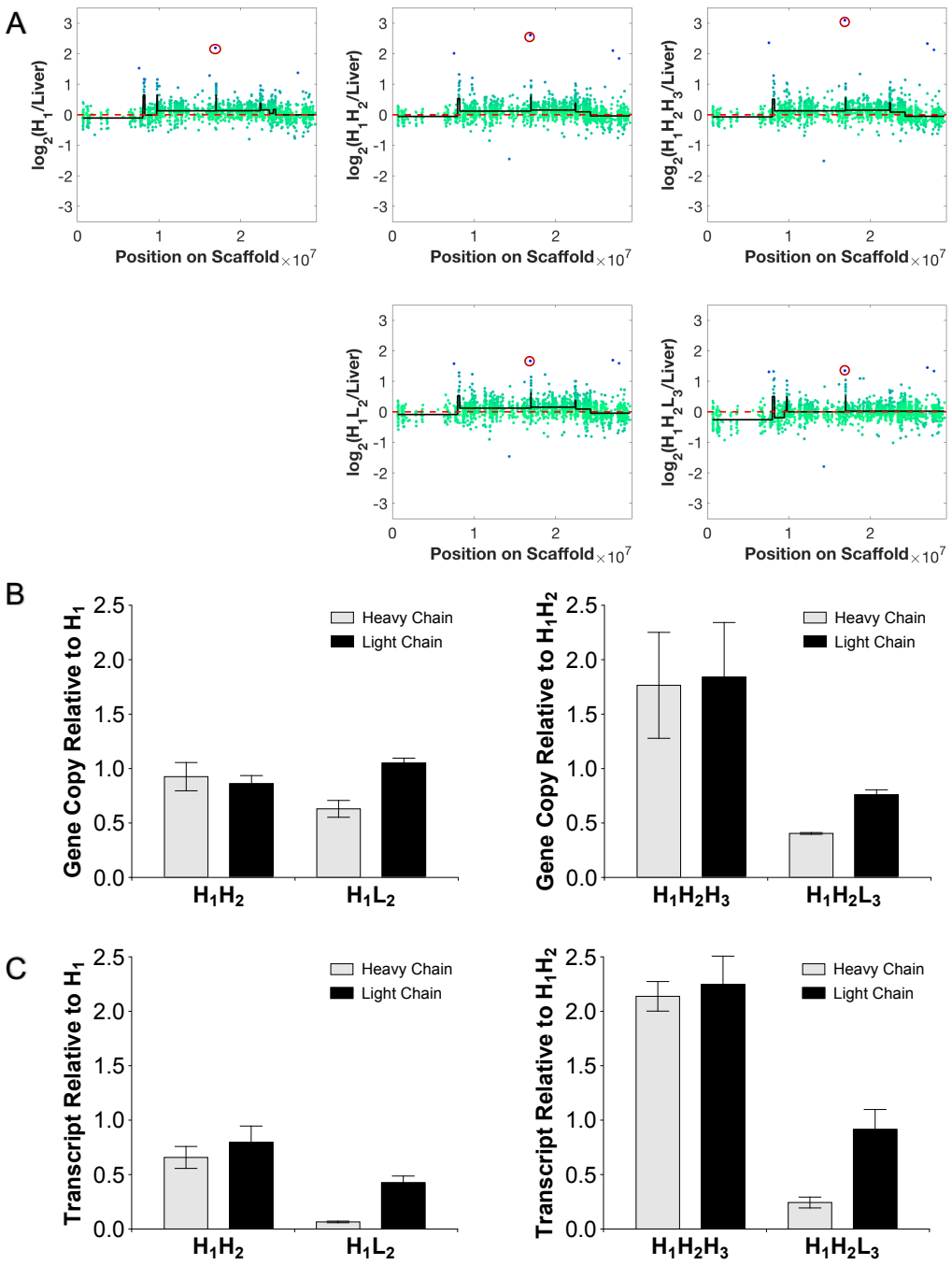

Figure 4-11. Loss of genome copy number at *Rc3h1* locus and as well as IgG gene copy number and transcript level in low producers. (A) Two independent losses of the gene probe for the integration site at *Rc3h1* (circled in red) seen during high-to-low producer transition. The green-blue dots represent the log-ratio intensities for each probe in the region relative to liver. The black line is the mean intensity of the segments as identified by the DNAcopy. (B) Genomic qPCR results show loss of heavy chain correlating with the loss of the integration site. Results reported in terms of gDNA copy number fold change relative to the subclone parent. (C) Transcript expression as quantified by qRT-PCR shows loss of heavy and light chain transcript levels. Results reported in terms of transcript expression fold change relative to the subclone parent.

This integration site was further investigated using the sequencing reads pileup of H_1 and the parental cell line DG44 (Figure 4-12). Figure 4-12A shows a 29 Mbp genomic region where the integration site is marked with a red vertical line. Though the read pileups for DG44 and H_1 look very similar across the entire region, the region marked in yellow shows a slight (~1.2-fold) amplification and the integration site shows about a 10-fold amplification. Zooming in further (Figure 4-12B), we see the integration locus is in a smaller segment of 31 Kbp that was amplified to ~8 times from diploid. DELLY identified tandem duplications and inversions within this segment, as well a large deletion. No deletions, duplications, or inversions were present at this site in DG44; there is thus no indication that the region surrounding this site is structurally unstable or has a higher propensity for duplication/amplification in the host cell line. The structural changes incurred after vector integration and amplification may have contributed to the instability of the locus and the loss in transgene copy number. Further zooming into the integration locus shows the sharp vector-genome boundary (Figure 4-12C).

The sequence pileup around integration sites #2 and #3 (Table 4-1) are shown in Figure 4-13A and Figure 4-14A, along with the CGH data in the integration regions. Site 2 had an amplification with respect to DG44 (Figure 4-13A), while site 3 did not show any amplification. The CGH data of the probes in the region did not show a significant difference in copy number among high and low producers (Figure 4-13B and Figure 4-14B).

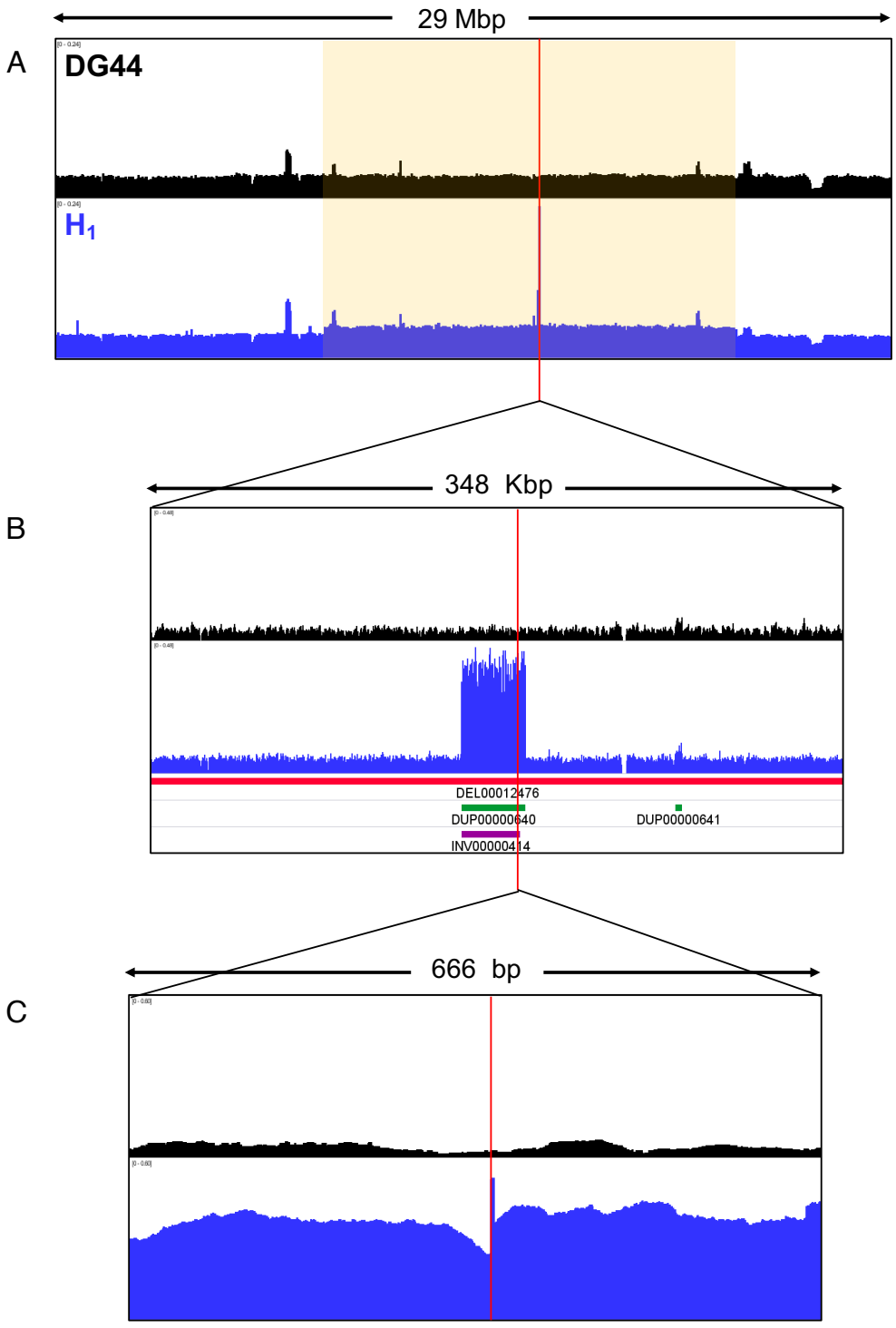

Figure 4-12. Normalized sequencing read pileup (Reads per Million) for DG44 and H_1 in the integration region (*Rc3h1*) shown at different magnifications. The red vertical line denotes the integration junction. The red horizontal bar indicates deletion, the green bar indicates tandem duplication, and the purple bar indicates inversion.

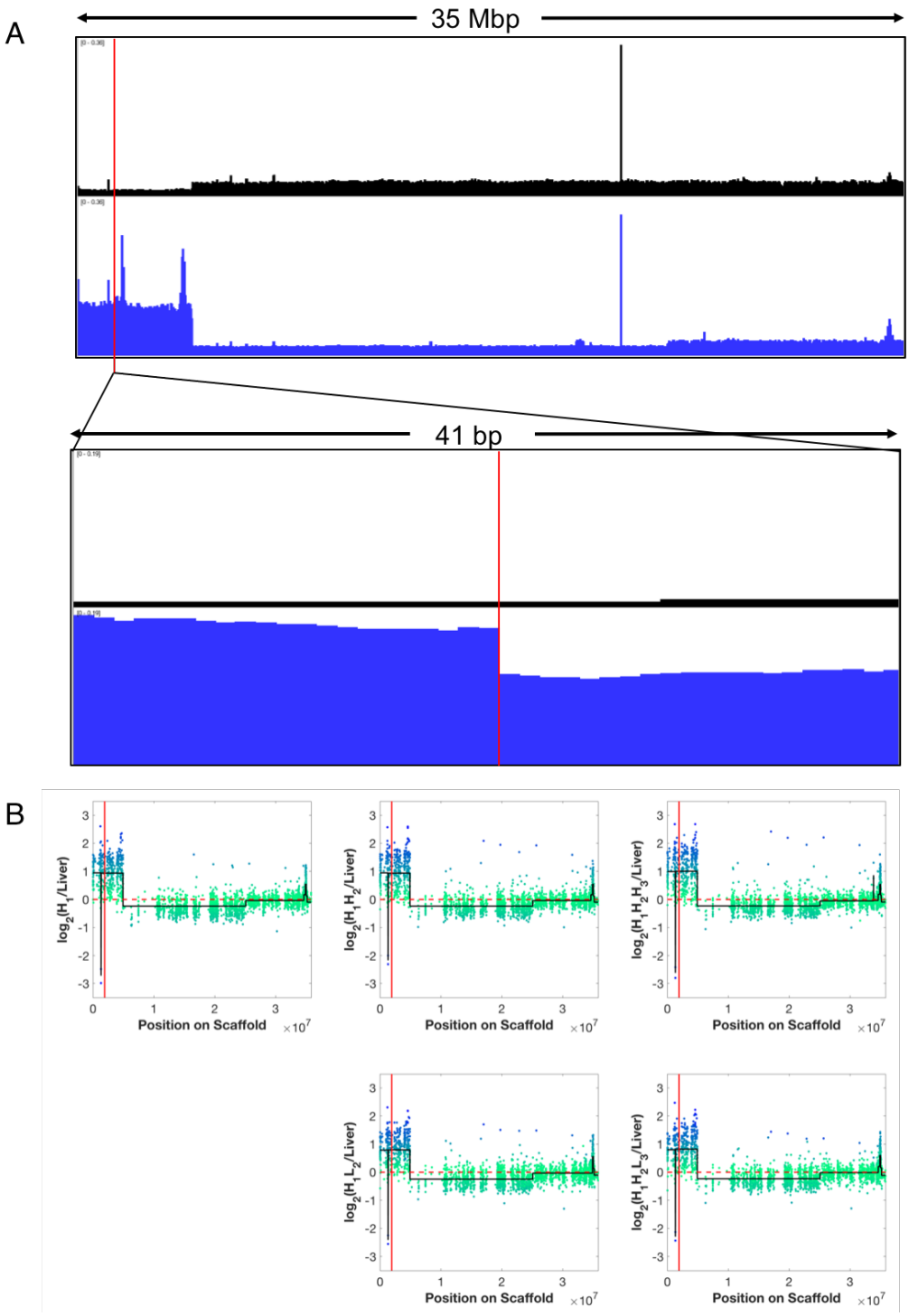

Figure 4-13. Genome copy number at integration site #2. (A) Normalized sequencing read pileup (Reads per Million) for DG44 and H_1 for integration site #2. The red vertical line denotes the integration junction. (B) CGH data for the five sublines show no significant change among high and low producing clones. The green-blue dots represent the log-ratio intensities each probe in the region relative to liver. The black line is the mean intensity of the segments as identified by the DNAcopy. The red vertical line denotes the integration junction.

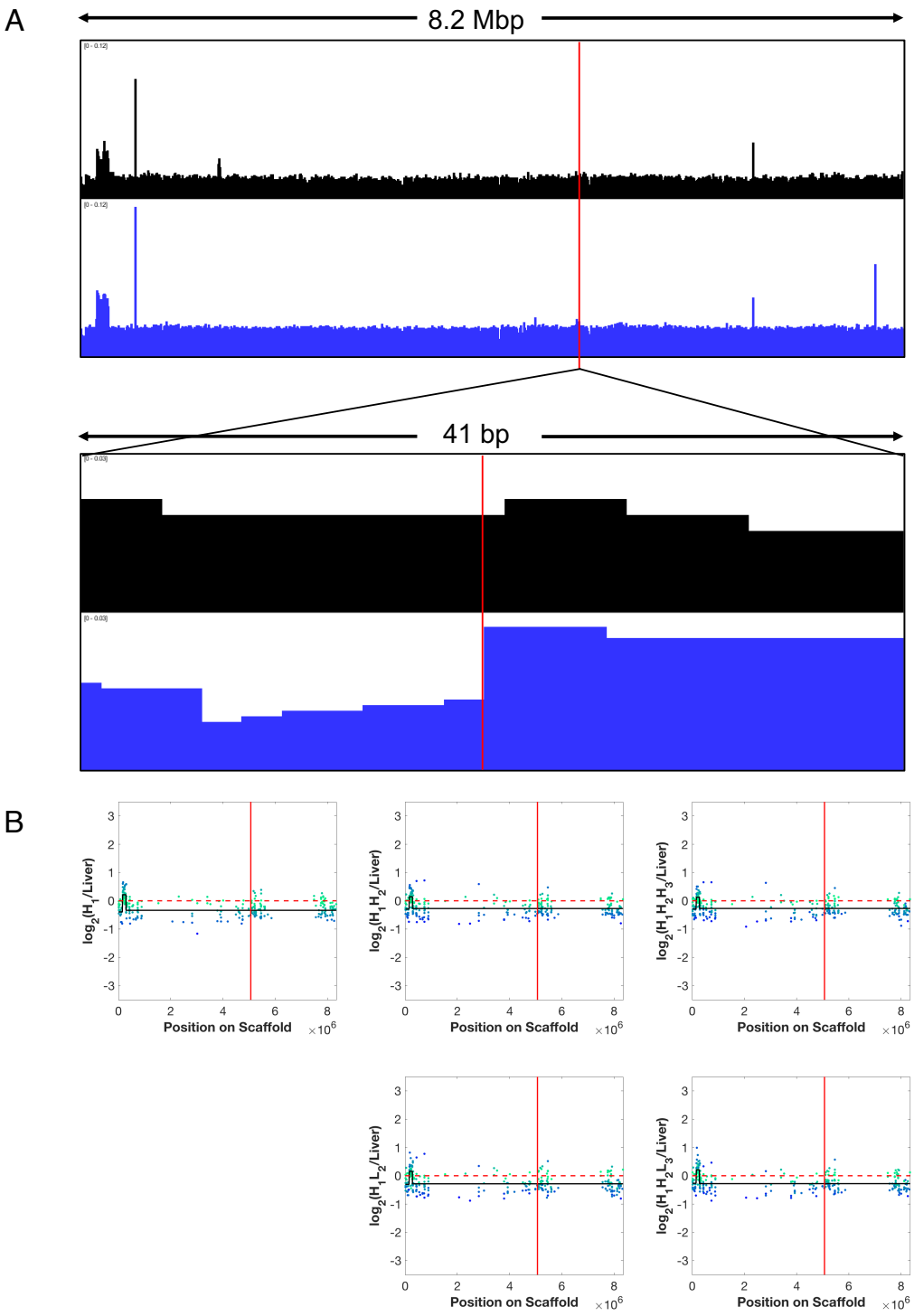

Figure 4-14. Genome copy number at integration site #3. (A) Normalized sequencing read pileup (Reads per Million) for DG44 and H_1 for integration site #3. The red vertical line denotes the integration junction. (B) CGH data for the five sublines show no significant change among high and low producing clones. The green-blue dots represent the log-ratio intensities each probe in the region relative to liver. The black line is the mean intensity of the segments as identified by the DNAcopy. The red vertical line denotes the integration junction.

4.5 Discussion

4.5.1 The stability of recombinant CHO cell lines

Cell lines used in industrial production of biopharmaceuticals were initially clonal, originating from a single cell at some stage after the introduction of the transgene into the genome. This is to avoid the risk of a subpopulation within a heterogeneous population eventually outgrowing the others and, as a result, altering the productivity or the product quality. It is now a regulatory requirement that production cell lines be initially phenotypically and genomically "clonal" [45, 62]. Recent studies have addressed the issue of the variability among single cell clones from clonally derived cell lines [75], and discussed the possibility for using cell pools for the production of non-clinical material [94].

Karyotyping of CHO host cells and assessment of chromosomal rearrangements has been demonstrated using various techniques such as G-banding, BAC-FISH (Bacterial Artificial Chromosome Fluorescence in Situ Hybridization), and chromosome painting [22, 23, 69, 71, 95-97]. DG44 was reported to have twenty chromosomes, with seven normal chromosomes, eleven chromosomes with rearrangements of normal chromosomes, and two uncharacterized marker chromosomes [95]. Other karyotyping studies on DG44 and CHO-K1 have found that chromosomes 1, 2, 4, 5, 9, and 10 (by classical nomenclature) are likely to have at least one intact copy [22, 23, 71, 98]. Our karyotyping data showed the presence of at least two copies of chromosomes 1 and 6, and at least one copy of chromosomes 2, 7, 8, 9, and 11. Comparing all of these studies, at least one intact copy of chromosomes 1, 2, and 9 seems to be stably present in all tested cell lines, and so these are promising targets for transgene integration.

Previous studies have also found a wide distribution of chromosome number in CHO cell lines, ranging from 15 to 50 chromosomes [23, 68, 69]. Derouazi et al. [71] found that the chromosome number for sixteen DG44-derived recombinant cell lines varied from 19 to 41. In our study of five cell lines also derived from DG44, we found a wide spread of chromosome numbers (20 - 70 chromosomes, Figure 4-4C). Importantly, we demonstrated that this distribution is rapidly re-established by about 30 generations after starting from a single cell.

4.5.2 Clonal cells and population heterogeneity

By single cell cloning to ensure that all offspring cells originated from a particular karyotype and genome structure, we showed chromosomal reorganization and genomic structural changes occurred again to give a population with a wide distribution in karyotype and chromosome number. Interestingly, this genomic restructuring was accompanied by relatively small levels of gain and loss of gene copies. The number of gene probes that showed gain or loss of copies from a diploid genome was about 4-5% in all the cell lines (Figure 4-6). Further, in each subcloning, over the span of 30-40 population doublings, approximately 100 to 2000 probes (0.06 to 1.16 % of all probes) gained or lost copies from their immediate parent. However, the CGH microarray that we employed did not have a high spatial resolution. The Chinese hamster genome assembly and the sequencing analytical tools also await further enhancement. We thus refrain from determining the fraction of total genome regions that are subject to gain or loss of copies. Nevertheless, the CGH and sequencing data both indicate that the genes that changed copy number in CHO cell lines constitute only a minor fraction.

It is also interesting to note that the number of gene probes that showed a loss of copy is relatively small compared to those which gained copy. It is plausible that a cell with a more drastic loss of copies may have a growth disadvantage or be prone to death. Similarly, the cells that have an "extreme" karyotype (too many, too few, or too drastic of a change (e.g. losing both alleles of critical regions)) may not survive. Therefore, even though cells continue to undergo genomic structural changes, the population still maintains a relatively stable distribution of karyotype and gene copies. Taken together, the cell lines established by subcloning largely returned to a distribution of gene copy and chromosomal organization that is very similar to their parent. However, in this study we examined only the heterogeneity in the chromosome number, gene copy number, and the productivity phenotype. Even though only a small fraction of genes changed copy number in each pair of parent-subclone, extensive sequential subcloning will increase the probability of generating a new population that has a skewed property distribution compared to the ancestor cell line.

This thus raises the prospect that, in the "recloning" or re-purifying a cell line, one should focus on maintaining the property distribution of the population of the production cell line. For example, pooling a large number of cells that maintain the productivity and eliminating low producers from the population, rather than picking a colony that grew out from a single cell.

4.5.3 Productivity stability and transgene integration

Traditionally, a CHO production cell line is generated by integrating the product transgene into a random locus in the genome followed by the amplification of transgene copy number. The initial integration of the transgene after transfection typically occurs at one or a few loci [71, 99, 100], and in the subsequent amplification the copy number increases over a wide range. The number of copies of transgene integrated is affected by the many factors, including the dose of plasmids and the extent of selective pressure applied. In some cases, a large number of copies were integrated even without amplification [54, 71, 101]. The amplification process typically amplifies not only the vector and transgene, but also the surrounding genomic region. Furthermore, breakage, ligation, and other structural changes may occur during amplification [78, 102, 103]. Loss of productivity in transgene amplified cell lines over long-term culture occurs frequently [78, 79, 104]. Hence, before a cell line is chosen to be the production cell line, extensive testing of its stability is commonly performed.

In this study, by identifying the transgene integration sites, we not only showed that transgene copy loss was likely the cause of the lost productivity, but also demonstrated that the lost copy in two independent isolations of low producing subclones was from the same integration site (*Rc3h1* locus). The loss was accompanied by a decreased IgG heavy chain copy number and transcript level. Sequence analysis of the production parental line (H_1) revealed extensive amplification and structural rearrangements in the *Rc3h1* locus (tandem duplication, inversion, and deletion), possibly occurring in conjunction with the amplification process during cell line development. The data thus suggests that some amplified sequences are unstable and prone to repeated loss of copy number.

Intriguingly, from CGH data we found another segment of an amplified 4 Mbp region in the production cell line was also lost in both low producing subclones (Figure 4-7A).

From our archived CGH data this segment showed frequent gain or loss of copies in other CHO cell lines (Figure 4-8). Although the segment did not harbor any integrated transgene, its loss appears to associate with the loss of the productivity or the loss of the transgene at the *Rc3h1* locus, as its copy number was lower only in the low producing subclones, but normal in the high producing subclones. It is possible that the region also harbors genes or epigenetic loci that confer traits related to productivity, although the region was not very transcriptionally active, and no notable gene was found. It is also possible the co-loss of copy number in this region and the *Rc3h1* locus was related to other clonal events. It is plausible that the low producing sublines H_1L_2 and $H_1H_2L_3$ arose from subclones that had DNA repair alterations that triggered copy loss in both *Rc3h1* and the 4 Mbp region, and these may be linked to transgene copy change and loss of productivity. Regardless of the mechanism of the co-loss of the two regions, the data suggests that some genome regions are possibly prone to structural variation. Our results also demonstrate that by combining CGH and subcloning one can identify such genome regions that are more vulnerable to copy number loss. One can thus use these tools to help determine clone stability.

With the advances in genome engineering, transgene integration can now be targeted to a specific locus in the host cell genome, without resorting to the traditional selective pressure-based transgene amplification [20, 105-107]. In general, the site selected for targeted integration should be stable, and not prone to structural variation or copy number changes. Genome scale CGH and sequencing analysis of structural variants provide valuable information on the regions to be avoided. Additionally, the target genomic regions should be transcriptionally active and epigenetically accessible for gene expression. Recent work from our laboratory showed that a single copy of the transgene integrated into a genomic region of high chromatin accessibility and high transcriptional activity can have an expression level equivalent to or higher than a cell line with multiple copies [18]. The incorporation of genomic analysis and genome engineering into cell line development will allow this traditionally empirical operation to become a design-based process.

5 Detecting common regions of structural variability across cell lines from whole genome sequencing

Reproduced from: O'Brien, C. M.*, O'Brien, S. A.*, Bandyopadhyay, A. A., Hu, W. S. Detecting common regions of structural variation across cell lines from whole genome sequencing. (In preparation) *CMO and SAO contributed equally to this work.

SAO processed the sequencing data. CMO and SAO performed data analysis and visualization of the results.

5.1 Introduction

Protein therapeutics, a type of biologic, are large pharmaceutical molecules typically produced within cells due to their size, difficulty in synthesis, and complexity of post-translational modification and protein folding. The most commonly used cell type for biologics production is the Chinese Hamster Ovary (CHO) cell, which is capable of human compatible protein modifications and high protein expression.

Typically, the sequence encoding the biologic is inserted into a vector and integrated into a cell line which can be used for long-term production of the protein. The traditional cell line development process entails random integration of the vector into the genome and subsequent selection to identify a cell clone which grows well and produces large quantities of protein. However, this random integration process is inherently variable – some integrations of the product gene are liable to be subject to genomic change, be it structural or epigenetic. This process has great potential for improvement through the use of targeted integration to drive repeatable cell line generation by selecting a desirable site for integration.

Targeted integration requires knowledge of the cell line and genome so that a site may be rationally selected. An ideal site would have both a stable genome as well as high transcriptional activity and lack of silencing. These characteristics may be identified through the use of whole genome sequencing, RNAseq, and epigenetic methods such as

ATAC-seq [18]. By combining these different types of information, a desirable site may be selected.

Choosing a poor integration site affects not only the short-term protein production level but can cause loss in productivity over the lifetime of the cell line due to genomic rearrangement or silencing of the product gene [79]. This can result in drift of the product over the product lifecycle and changes to titer, cell metabolism, and even product quality. Stability is thus a highly desirable characteristic for a production cell line.

One way to assess genomic stability is the use of structural variants (SV), which consist of larger rearrangements of the genome, including deletions, duplications, insertions, translocations, and inversions. These types of genomic changes have the potential to alter product gene expression. Detection of SVs has been used to identify characteristics of different cancerous tumors [108], due to the recurring changes seen within different types of cancer. These methods identify regions in the genome with more or less frequent changes, which may serve as a hallmark for local genomic stability for a given cell type.

In this work, we use the SV-detecting tools DELLY and LUMPY on whole genome sequencing data of 23 cell lines to assess what kinds of structural vulnerabilities exist within and across cell lines. This data may be used as a tool to identify potentially favorable integration sites by looking for those regions with little to no genomic rearrangements, indicating that these regions may be less prone to copy loss of an integrated transgene, and would thus provide long-term, stable expression.

5.2 Materials and Methods

5.2.1 Sequencing data

Whole genome sequencing data from 23 CHO cell lines and Chinese hamster tissue was used for this study. A list of samples, the number of reads for each sample, and their source is listed in Table 5-1. In total, 11 CHO-K1 based cell lines, 7 CHO-S based cell lines, 2 DG44 based cell lines, and 3 DXB11 based cell lines were included in the analysis. These consisted of a mix of host and production cell lines. Data consisted of paired end reads ranging from 82 to 250 bp in length.

Table 5-1. List of whole genome sequencing samples used for this study.

Cell lineage	Cell line	Type	# of Reads	Source (SRA when applicable)
CHO-K1	CHO-K1 ECACC	Host	182,320,492	SRR803176 - SRR803178 [84]
	CHO-K1 SF	Host	175,891,332	SRR803179 - SRR803181 [84]
	CHO-K1 PF	Host	176,348,036	SRR803173 - SRR803175 [84]
	CHO-K1 ATCC	Host	827,770,910	SRR329939 - SRR329954 [109]
	CHOK1-A	Host	504,704,636	Internal Data
	rCHOK1-A1	Production	321,048,726	Internal Data
	rCHOK1-A2	Production	332,061,589	Internal Data
	rCHOK1-B1	Production	466,464,265	Internal Data
	rCHOK1-B2	Production	479,013,314	Internal Data
	rCHOK1-B3	Production	467,925,056	Internal Data
	SH87	Production	667,180,442	SRR5378587, SRR5378591, SRR5378596, SRR5378598 - SRR5378604 [54]
CHO-S	CHO-S	Host	146,422,084	SRR803182, SRR803183 [84]
	C0101	Production	397,200,133	SRR801491 - SRR801496 [84]
	rCHOS-1	Production	187,651,305	Internal Data
	rCHOS-2	Production	117,090,818	Internal Data
	rCHOS-3	Production	136,833,174	Internal Data
	rCHOS-4	Production	136,561,351	Internal Data
	rCHOS-5	Production	169,884,959	Internal Data
DG44	DG44	Host	149,416,821	SRR803184, SRR803185 [84]
	rDG-IgG	Production	898,046,557	Internal Data
DXB11	DXB11	Host	381,195,209	SRR1561427, SRR1561428, SRR1561441, SRR1561442 [110]
	rDXB11-1	Production	117,846,515	Internal Data
	rDXB11-2	Production	150,070,758	Internal Data
N/A	Chinese Hamster Tissue	N/A	507,430,284	SRR954911- SRR954915 [84]

5.2.2 Data Processing

Pre-processing

An overview of the processing pipeline is in Figure 5-1. Raw sequencing data was trimmed to remove sequencing adapters using Trimmomatic [55] (version 0.33). Paired reads were then mapped to the CriGri-PICR release of the Chinese hamster genome [56] using BWA-MEM [57] (BWA release 0.7.17). Reads with a MAPQ score lower than 20

were removed using Samtools [58] (version 1.9). Duplicates reads were then marked using the MarkDuplicates command in Picard tools 2.18.16.

Structural Variant detection

Structural variant detection was performed on duplicate marked, quality filtered BAM files using two different software packages, DELLY [88] (version 0.8.1) and LUMPY [111] (lumpyexpress, version 0.2.13). Structural variants from both packages were filtered to obtain precise variants calls with at least 5 split reads supporting them. Filtered variants from DELLY and LUMPY were merged using bedtools [112] (version 2.29.2). To avoid duplicate variant calls spanning almost identical regions, deletions or duplications with positions within 10 bp on each end were merged using MATLAB. Further analysis of structural variants and vulnerable regions was performed using bedtools.

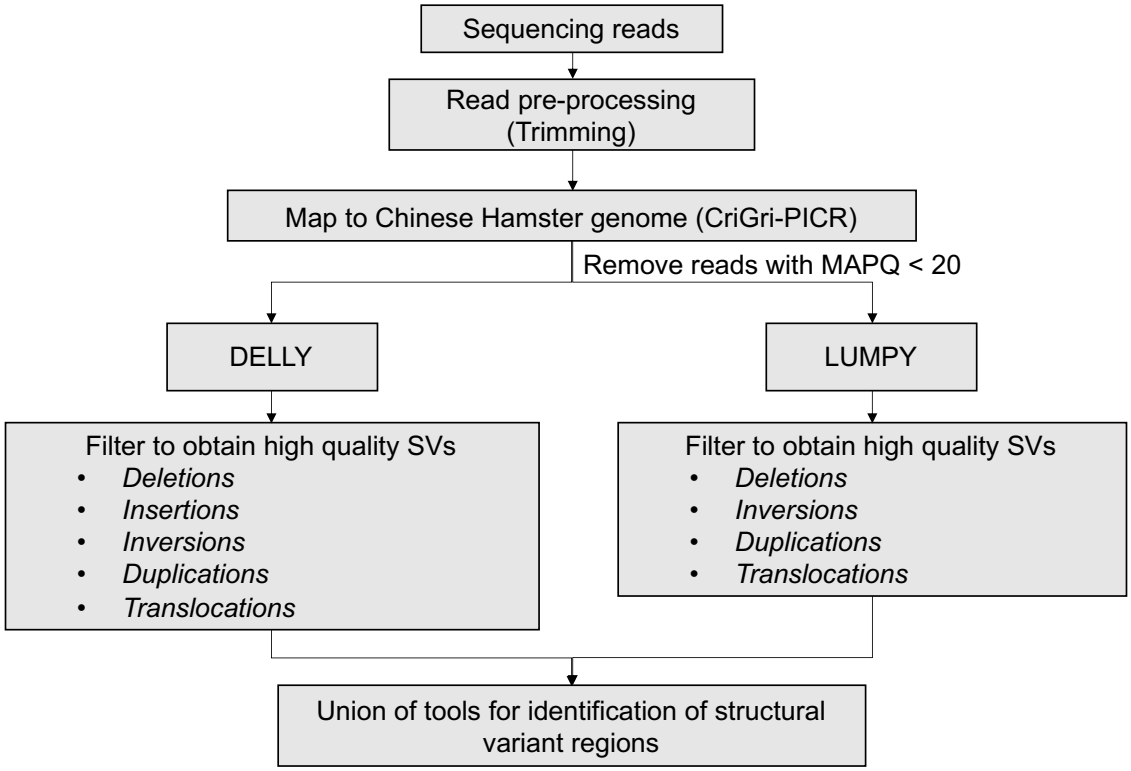

Figure 5-1. Overview of data processing pipeline.

5.2.3 Statistical Analysis

The non-parametric Mann-Whitney Test for non-normal distributions was used for statistical analysis in Mathematica.

5.3 Results and Discussion

5.3.1 Types of structural variants in different cell lines

We first processed the raw data from the 23 sequenced CHO cells lines following the procedure outlined in the methods, identifying SVs using the packages DELLY and LUMPY. The combination of these packages was chosen given the difference in detection methods, such that merging the identified SVs from these two tools and filtering to remove duplicates led to a more complete dataset. Additionally, genome sequencing from Chinese Hamster tissue was used as a control for false positives. In this work, deletions, insertions, inversions, duplications, and translocations were identified for analysis. The number of variant calls of each type are shown in Figure 5-2. The largest class of variants were deletions, followed by translocations, duplications, inversions, and then insertions.

Overall, production cell lines appeared to have more variant calls, especially those derived from CHO-K1. This could be due to the high stress the cells endure during the selection process. Additionally, cells from similar backgrounds tend to have similar numbers of variants. CHOK1-ECACC and the two cell lines derived from it (CHOK1-PF and CHOK1-SF) have a much lower number of variant calls than the remaining CHO-K1 cell lines derived from CHOK1-ATCC. This suggests many structural variants may be lineage dependent. Alternatively, the number of variants could be related to read depth for each sample, but this does not seem to be the case as some cell lines with more reads had fewer variant calls, such as DXB11 having fewer variant calls than the two DXB11 production cell lines rDXB11-1 and -2. The hamster tissue sample had a fair number of structural variants reported. This could be due to genome mis-assembly, sequencing error, error in library creation, or difficulties in mapping. However, this information can be used to correct for false positives in other cell lines.

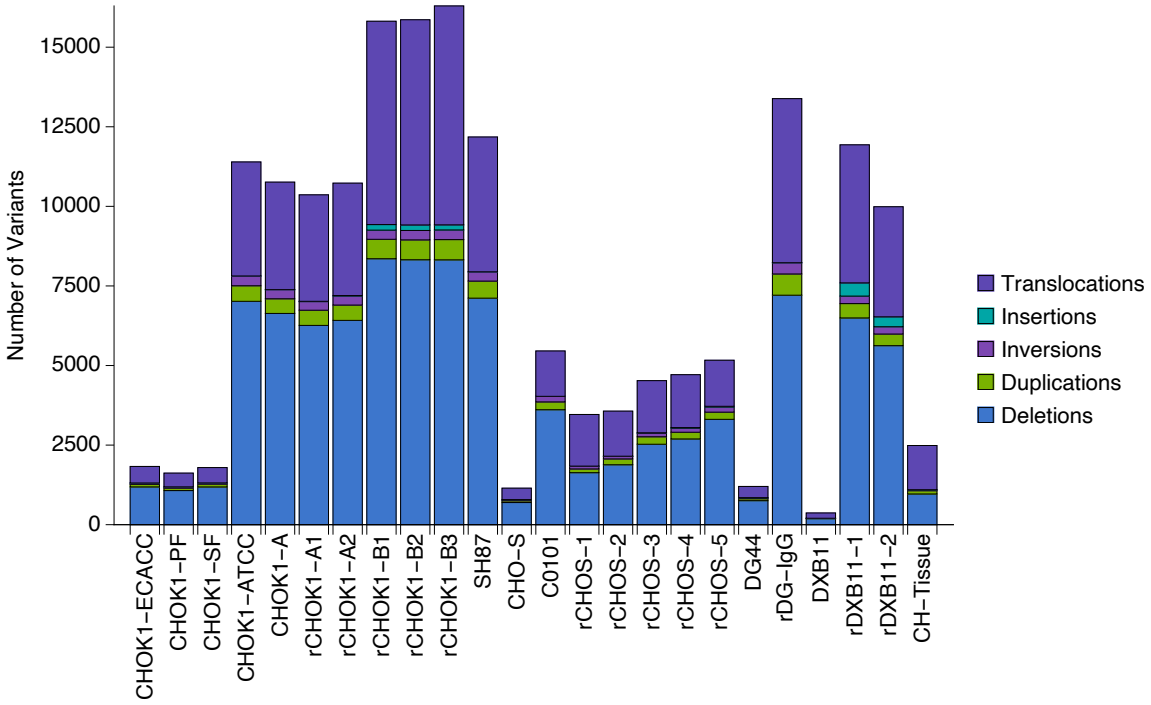

Figure 5-2. Number of structural variants of each type for the 23 cell lines examined.

For the remainder of this work, we chose to focus on deletions and duplications, as there were very few insertions and inversions, while translocations are more challenging to interpret for understanding genomic instability.

5.3.2 Locations of SVs within the genome

To better understand the types of regions vulnerable to structural variation, we examined genomic features that were prone to variation, such as exons, introns, and intergenic regions. For this analysis, the fraction of a given genomic feature that was either deleted or duplicated was calculated. For example, the fraction of exons with deletion calls would be calculated as:

$$frac_{del} = \frac{\text{\# of bp with deletions that are in exons}}{\text{Total size of exon regions (bp)}}$$

The resulting data was visualized as a heat map, with a low fraction a gene feature having structural variants in purple, and a high fraction with structural variants in red (Figure 5-3A and B). Looking at deletions (Figure 5-3A), it seems that within most cell lines, no one gene feature is overrepresented among structurally variable DNA. This would be expected if the probability of a structural variant occurring is equal between all locations

in the genome. A few cell lines, such as rCHOK1-B1,2,3 and rDG-IgG, appear to have an increased density of deletions in pseudogene introns, though the implications of this are unclear. Between cell lines, there is again a great deal of variability. Corresponding to the analysis on total number of structural variants, host cells also have a lower fraction of variants in the different regions compared to production cell lines.

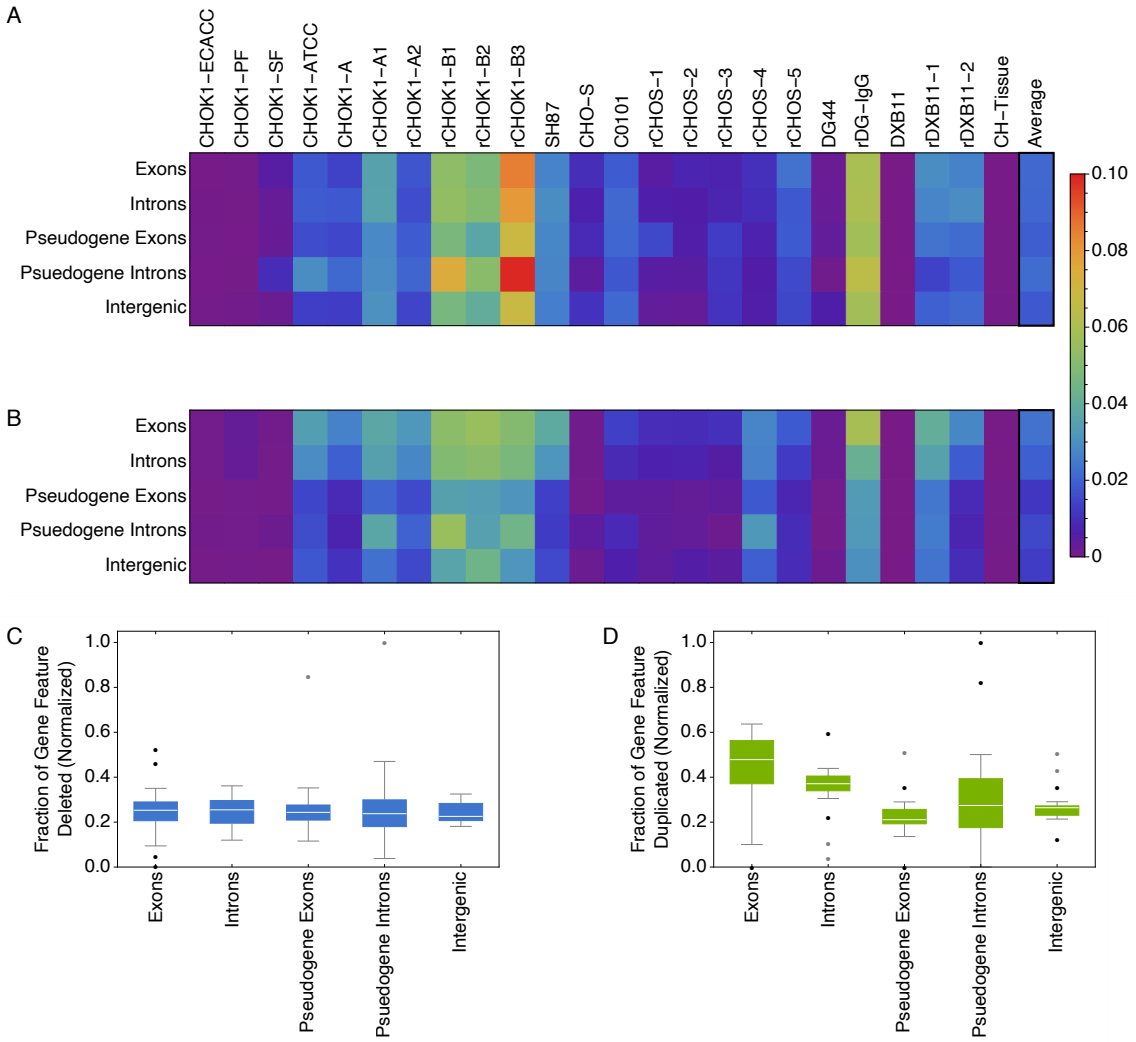

Figure 5-3. Fraction of gene features containing structural variants for each cell line. (A, B) Fraction of gene/pseudogene exons and introns or intergenic regions containing deletions (A) or duplications (B) for each cell line. Average across all cell lines is shown in the last column. (C, D) Box-whisker plot of fraction of gene feature that is deleted (C) or duplicated (D) across all cell lines, normalized to number of deletions or duplications in the cell line and rescaled from 0 to 1. White line represents the mean, and the box spans the 25% to 75% quantile. Outliers are shown as black dots and far outliers are gray dots.

For duplications, there appears to be more of a difference in the fraction of duplications between different types of gene features within some cell lines (Figure 5-3B). To further quantify these results, we examined the distribution of the fraction of each gene feature with a deletion or duplication, after normalizing to the number of total variants of that type found in each cell line to prevent skewing of the data towards cell lines with low numbers of variants. The data was further rescaled between 0 and 1 for visualization (Figure 5-3C and D). For deletions, there was no difference between the fraction deleted of different gene features, as expected from the heat map (Figure 5-3C). For duplications however, a significantly larger fraction of exons had duplications than all other gene features (Mann-Whitney test, $p < 0.005$ all comparisons) (Figure 5-3D). Introns were also significantly overrepresented relative to pseudogene exons/introns and intergenic regions (Mann-Whitney test, $p < 0.05$ all comparisons). This implies that duplications in these cell lines may be related to the expression of genes that confer a growth or other advantage.

5.3.3 SVs within a cell line to identify potential vulnerability

To further refine what is considered to be a "vulnerable region," we considered two different types of variability: intra- and inter-cell line.

To assess recurrent SVs within a cell line, we identify repeated deletions or duplications (schematic in Figure 5-4A). For this analysis, we considered locations which had two or more variants of at least 500bp. The presence of multiple, repeated SVs at a single locus within a single cell line indicates changes to either different chromosomes in one cell, or recurring changes within different subsets of the population. In either case, these repeated SVs may be indicative of genomic vulnerability.

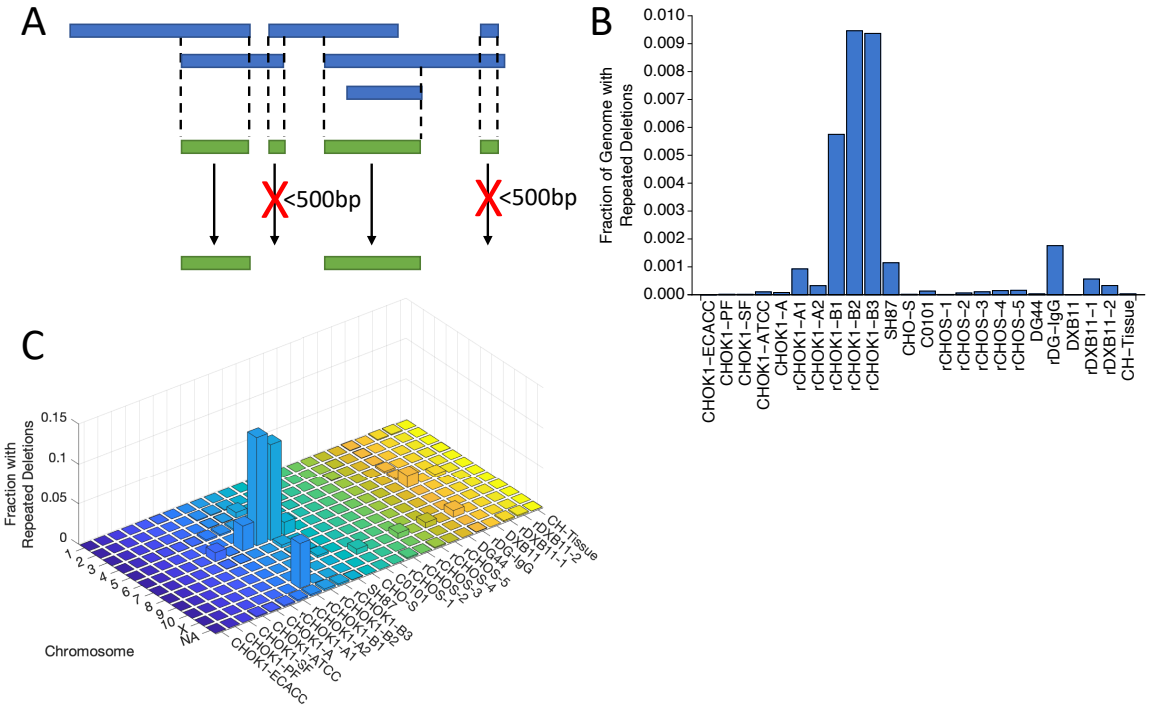

Figure 5-4. Analysis of repeated deletions within each cell line. (A) Detection of repeated deletions within one cell line. Overlapping variant calls of at least 500bp are collected for analysis. (B) Fraction of genome containing repeated deletions in each cell line. (C) Fraction of each chromosome containing repeated deletions for each cell line.

The fraction of each cell line with overlapping deletions is shown in Figure 5-4B. While these overlapping SVs comprise only on average a few percent of the whole genome, it does indicate regions of the genome particularly vulnerable to structural variation and which pose potentially a higher risk for transgene integration, as the gene may be subject to greater genomic change over the lifecycle of the cell line. This analysis was extended to examine if certain chromosomes have more vulnerable regions than others (Figure 5-4C). Overall, there was more variability between cell lines than between chromosomes, but chromosome 6 had a higher proportion of repeated deletions within multiple cell lines. This could be due to instability, or if these cell lines had many copies of chromosome 6, which could lead to different types of structural variants in the same location. CHO cells are aneuploid and the copy number of different chromosomes can vary based on the host cell [22, 25], which would explain potential chromosomal differences between different cell lines.

5.3.4 Sliding windows to identify common regions of vulnerability

Next, sliding windows were used to compute the average coverage of deletions and duplications across different regions in the genome. 100 kbp sliding windows with 75% overlap were used for this analysis (schematic in Figure 5-5A). This size was chosen as some enhancer elements are known to interact with DNA up to 50 kbp away [113], and so we can use this analysis to assess general regions for structural variability in the context of the broader genomic landscape.

Similarities in SVs between different cell lines were assessed by comparing the fraction of each window containing deletions or duplications across the genome. Hierarchical clustering using a Euclidean distance metric was used to look for cell lines with common regions of variation. Clustering the cell lines based on deletions showed that cell lines from the same lineage generally clustered together (Figure 5-5B), implying that some structural variation is inherited from the parental cell, although this was not always the case. For example, rDG-IgG had a high distance from its parental cell DG44, although this may be due to the large changes the cell line underwent during development, as amplification of the transgene was performed. For the fraction of each window duplicated, the cells did not cluster as well by lineage (Figure 5-5C). This may be due to the overall lower number of duplications, and that duplications may be biased for certain types of gene regions (Figure 5-3D).

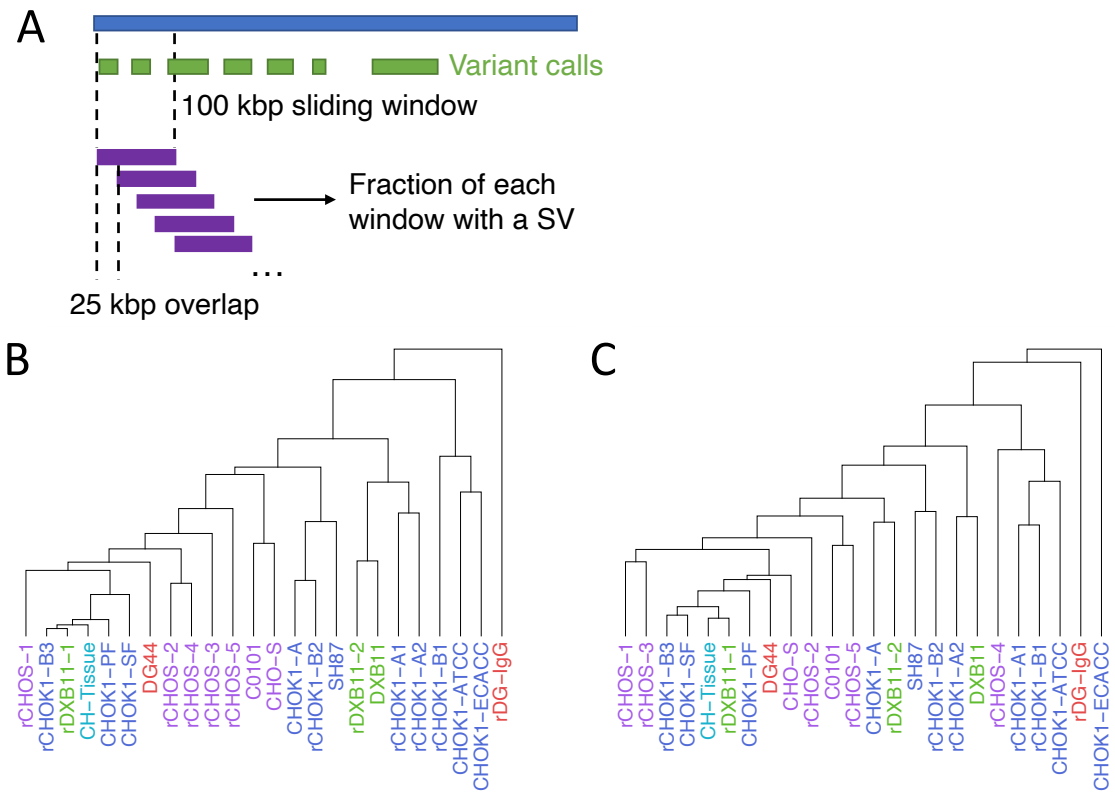

Figure 5-5. Window-based analysis of structural variation in different cell lines. (A) Sliding window analysis of variant calls. Fraction of each window covered by a structural variant is calculated. (B, C) Hierarchical clustering of cell lines based on fraction of genome window deleted (B) or duplicated (C). Cell lines are colored by host cell lineage.

The occurrence of deletions across the genome, averaged by cell lineage, is depicted in Figure 5-6. Here, Chinese Hamster liver tissue serves as a control to identify false positives on a genome-wide scale. Overall, the false positives are at a low frequency across the chromosomes in the genome, but at a higher rate in the scaffolds unassigned to chromosomes (Figure 5-6, labeled NA). Across the cell lines used in this study, there emerge regions of the genome which have significant overlap, indicating persistent SVs across all cell lines, and other regions which show little variation in any of the cell lines, indicating regions which are potentially more stable and may be more suitable for transgene integration from a stability perspective.

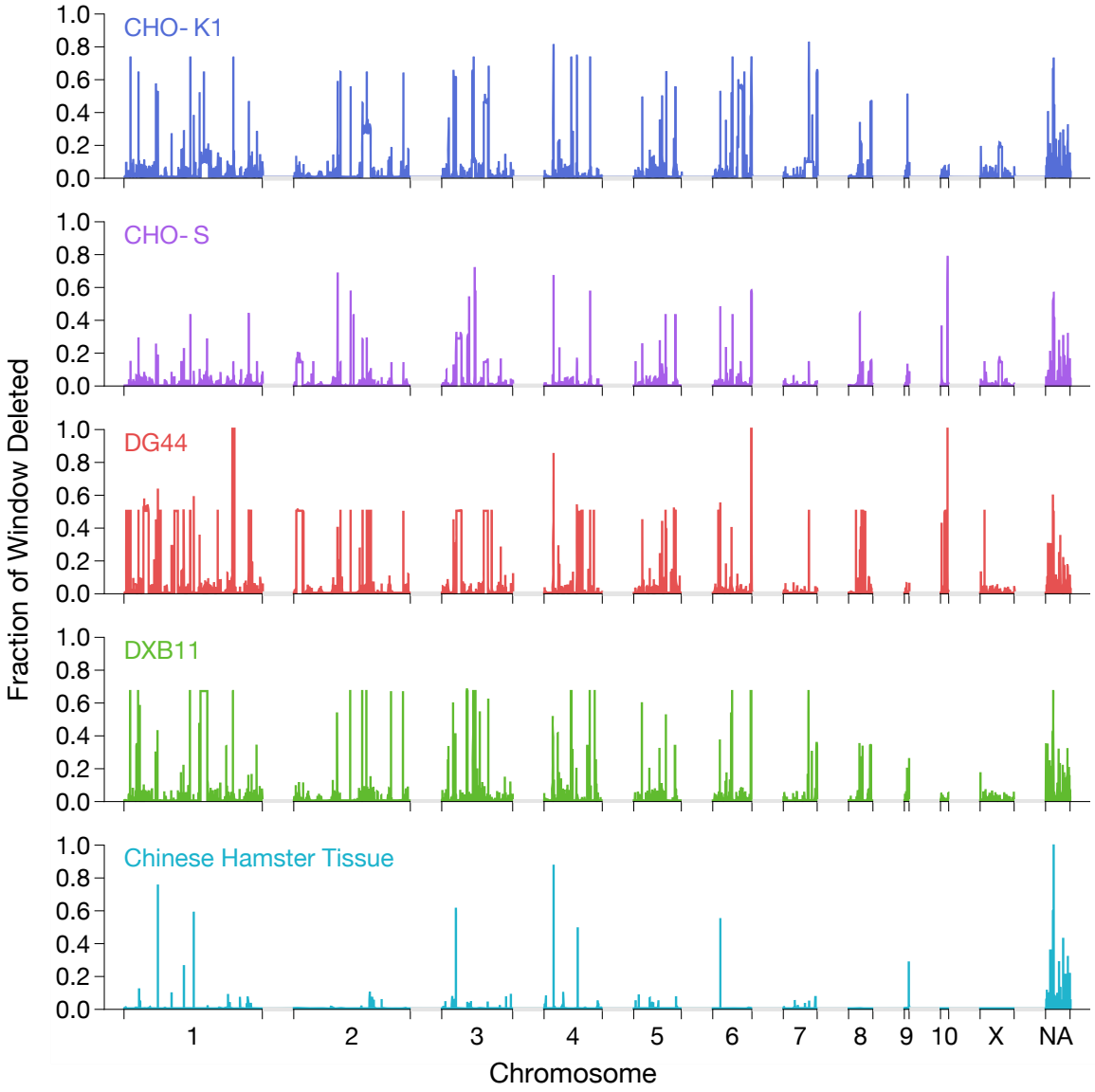

Figure 5-6. Analysis of structural variation on a genome scale. All scaffolds belonging to the same chromosome were stitched together, and the average fraction of a window that is deleted for a given parent cell is shown. Lines are colored by parental lineage.

To further examine regions of the genome vulnerable to change, we created genome tracks for visualization in IGV. All called deletions or duplications in the 23 CHO cell lines examined were intersected, and a count of the number of cell lines with a deletion or duplication at a particular genome locus was converted to an IGV track. Scaffold NW_020822439.1 is shown in Figure 5-7 as an example. Deletions and duplications each have their own track, and the presence of deletion or duplication calls in Chinese Hamster Tissue are shown in separate tracks as an indicator of false positives. In this scaffold, there

are clearly regions that tend to be deleted or duplicated in multiple cell lines, while other regions have no indication of structural variability. Additionally, some regions are called as deleted in Chinese Hamster Tissue, and so the deletion calls in those areas may be false positives. Using these tracks, a better understanding of structural variability in CHO cells can be gained.

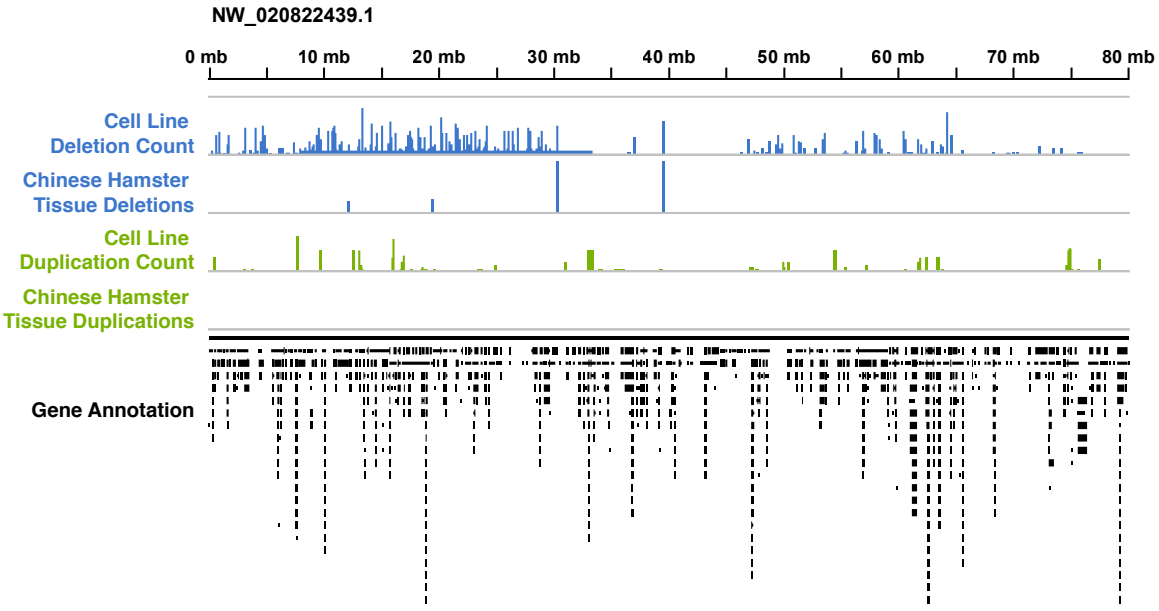

Figure 5-7. IGV tracks for visualization of deletions and duplications in the genome. Image of scaffold NW_020922439.1 on Chromosome 2. For Cell Line Deletion/Duplication Count, the bars show the number of cell lines with a deletion/duplication in that location, which can range from 0 to 23. Chinese Hamster Tissue Deletions/Duplications indicate the presence of a deletion/duplication detected from sequencing of Chinese Hamster Tissue in that location. The gene annotation track shows the locations of genes within this scaffold. Position in the scaffold is shown at the top of the figure.

5.4 Conclusion

In this paper, we have presented an analysis of genomic structural variation across 23 CHO cell lines to detect commonalities both within cell lines, and across different cell lines. While significant portions of the genome contained detected SVs, some regions and chromosomes had lower variability, and thus potentially more genomic stability, a desirable characteristic for a transgene integration site. By using the combined genomic tracks, we believe this work can serve as a guide in conjunction with information about transcriptional activity and epigenetic regulation to identify desirable sites for the targeted integration of transgenes for next generation cell line development.

6 Single copy transgene integration in a transcriptionally active site for recombinant protein synthesis

Reproduced with permission from: O'Brien, S. A., Lee, K., Fu, H. Y., Lee, Z., Le, T. S., Stach, C. S., McCann, M. G., Zhang, A. Q., Smanski, M. J., Somia, N. V., Hu, W. S. (2018). Single Copy Transgene Integration in a Transcriptionally Active Site for Recombinant Protein Synthesis. *Biotechnol J, 13*(10), e1800226. doi:10.1002/biot.201800226

6.1 Summary

For the biomanufacturing of protein biologics, establishing stable cell lines with high transgene transcription is critical for high productivity. Modern genome engineering tools can direct transgene insertion to a specified genomic locus and can potentially become a valuable tool for cell line generation. In this study, we surveyed transgene integration sites and their transcriptional activity to identify characteristics of desirable regions. A lentivirus containing destabilized Green Fluorescent Protein (dGFP) was used to infect Chinese hamster ovary cells at a low multiplicity of infection, and cells with high or low GFP fluorescence were isolated. RNA sequencing and Assay for Transposase Accessible Chromatin using sequencing data showed integration sites with high GFP expression are in larger regions of high transcriptional activity and accessibility, but not necessarily within highly transcribed genes. This method was used to obtain high Immunoglobulin G (IgG) expressing cell lines with a single copy of the transgene integrated into transcriptionally active and accessible genomic regions. Dual recombinase-mediated cassette exchange was then employed to swap the IgG transgene for erythropoietin or tumor necrosis factor receptor-Fc. This work thus highlights a strategy to identify desirable sites for transgene integration and to streamline the development of new product producing cell lines.

6.2 Introduction

Protein therapeutics have transformed the treatment of many complex diseases. In the past five years, over five dozen recombinant therapeutic proteins were approved by the FDA [114]. These new products include monoclonal antibodies directed at PD-1 (Keytruda (Merck) and Opdivo (Bristol Meyers Squibb)) and PD-L1 (Tecentriq; Genentech) for treating melanoma and bladder cancers, Fc fusion protein of Factor VIII/IX for hemophilia (Eloctate and Alprolix; Biogen), and bispecific antibodies anti-CD19/anti-CD3 for the treatment of leukemia (Blincyto; Amgen). Many more biologics are currently under development and in late phase clinical trials. With this new generation of biologics, it is increasingly important to have efficient methods for their production.

Chinese Hamster Ovary (CHO) cells are the most commonly used host cells for the manufacture of recombinant therapeutics [44, 115]. Traditionally, production cell lines are generated by random integration of the product transgene into the host cell followed by gene amplification and screening for high producing cell clones. Although this method of cell line generation has been in practice for over three decades and such cell lines have been used to produce the majority of the products on the market, the success of the method relies on screening a large number of clones for their productivity and stability over time [79, 116-119]. With the advances in genome engineering, new ways of generating cell lines, especially directing the integration of transgenes into specific loci in the host cell genome, are being explored.

There are a variety of tools available for the targeted integration of transgenes into specific loci, such as zinc finger nucleases, TALE nucleases, and CRISPR/Cas9 [120-122]. Another possible strategy called Recombinase Mediated Cassette Exchange (RMCE) allows for replacement of the sequence between two non-compatible recombination sites, previously inserted into the genome, with a new gene. A number of methods have been developed for RMCE using site specific recombinases, such as the Cre/*lox,* Flp/*FRT,* and ΦC31/*att* P/-B systems [123]. Integration of transgenes into CHO cells has been achieved at a reasonable efficiency using both targeted integration [20, 105, 107] and cassette exchange [19, 124-126]. Another similar method, dual RMCE, utilizes two different recombinases (such as Flp and Cre). Dual RMCE was reported to have a higher integration

efficiency than traditional RMCE by preventing excision or other internal recombination effects [127, 128], and has been used efficiently in CHO [129]. The success of these methods hinges on prior knowledge of a desirable site for transgene integration. One approach is to select a site identified from a high producing cell line. However, a high producing cell line generated using past methods often has multiple transgenes integrated into the host cell genome, and identifying the site that contributes most to transgene transcription is time consuming, often involving serial knockout of the multiple transgene copies.

The integration site of a transgene has an effect on its expression. Insertion of a transgene into locations with known transcriptional activity has been suggested to increase its expression [130, 131]. Additionally, expression of reporter genes at a particular site has been shown to be affected by proximity to active genes [132]. Using PiggyBac transposons or HIV-1 based lentivectors, transgenes have been preferentially inserted into transcriptionally active genomic regions [133-136]. These methods have been used for product transgene integration in CHO cells to obtain high producing cell lines [17, 137, 138]. The transcriptional activity of a gene is also affected by regional chromatin organization and accessibility [139-141]. Recently, a decrease of transgene productivity in CHO cells was reported to correlate with increased nucleosome occupancy [82].

Transcriptional activation has been shown to be associated with regions of the genome lacking nucleosomes, which are often referred to as open or accessible chromatin [140]. A number of methods are commonly used to assess the openness of chromatin structure, including MNase-seq [142], DNAse-seq [143], and FAIRE-seq [144]. Assay for Transposase Accessible Chromatin using sequencing (ATAC-seq) has recently emerged as an easily accessible method for genome wide evaluation of chromatin accessibility [145]. In this technique, a hyperactive Tn5 transposase tags native chromatin from cellular nuclei with high-throughput sequencing adapters, simultaneously fragmenting the DNA. After sequencing of the digested fragments and mapping to the genome, the frequency of the appearance of a particular sequence is indicative of the accessibility of the corresponding genomic locus [146]. Regions low in ATAC-seq signals are inferred as being occupied by nucleosomes [140].

In this study, we used a lentiviral vector with a reporter gene to evaluate the relationship between transgene expression and the local context of the genome in terms of transcriptional activity and accessibility. We further demonstrate the establishment of Immunoglobulin G (IgG) producing cell lines with a single integration site in regions with high transcriptional activity and open chromatin. Finally, we utilized dual RMCE by replacing the IgG transgene with two different product genes, streamlining the creation of new product producing cell lines.

6.3 Experimental Section

6.3.1 Vector Construction

The lentivectors pLOVE-dGFP and pLOVE-IgG-dGFP were generated using pLOVE empty vector (addgene Plasmid #15948) as a backbone. pLOVE empty vector was digested with PacI and AscI to insert either destabilized GFP (dGFP), or IgG-IRES-dGFP flanked by recombination sites (Figure 6-1A and B). The IgG-IRES-dGFP insert contains IgG heavy chain and light chain linked by a porcine teschovirus-1 2A (P2A) linker peptide [147], followed by an internal ribosome entry site (IRES) to drive translation of dGFP. The IgG is flanked by Lox511 and a minimal 34 bp flippase recognition target (FRT) site [148].

Donor vectors Lox511-PuroR-P2A-mCherry-P2A-TNFR-Fc-FRT and Lox511-PuroR-P2A-mCherry-P2A-EPO-FRT were constructed using Gibson assembly [149] to combine the different elements with a minimal backbone containing a Kanamycin resistance gene for selection in bacteria and a ColE1 origin of replication. The vector is promoter-less, and contains Puromycin resistance, mCherry (a red fluorescent protein), and a product gene (either Entanercept (a TNFR-Fc fusion protein) or Darbepoetin alfa (an engineered variant of Erythropoietin (EPO)), all linked via P2A, and flanked by Lox511 and a minimal FRT site. A schematic of the plasmid map is shown in Figure 6-1C.

Recombinase vector F2AC contains flippase (Flpe) linked to cre recombinase via a Thoseaasigna virus 2A (T2A) linker and driven by a CMV promoter [129].

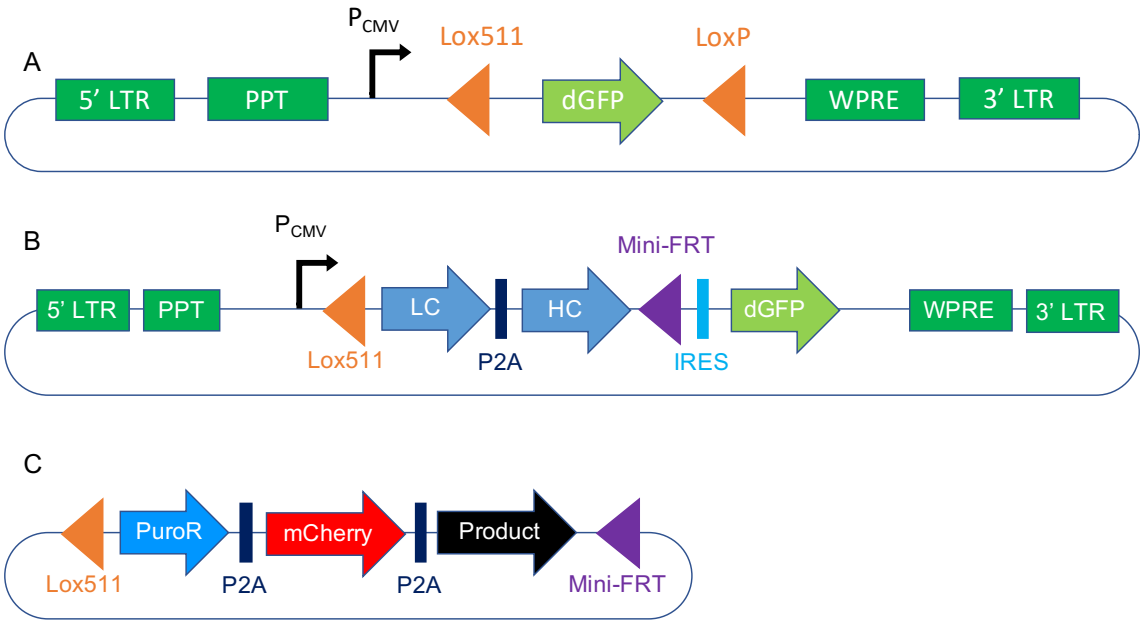

Figure 6-1. Vector schematics. (A) pLOVE dGFP lentivector. (B) pLOVE IgG-IRES-dGFP lentivector. (C) RMCE Swapping Vector.

6.3.2 Cell Culture

CHO-K1 cells (ATCC CCL-61, Manassas, VA) and derived subclones were cultured in F12K Medium (Kaighn's Modification, Gibco, Waltham, MA) supplemented with 10% Fetal Bovine Serum (Gibco, Waltham, MA) and incubated at 37°C in 5% CO_2. Single cell cloning was performed via limiting dilution. Cells were diluted to 1 cell per well (100µL) and plated into tissue culture treated 96-well plates.

6.3.3 Lentivirus production and infection

Lentiviruses were produced in 293T cells by co-transfection of pLOVE derived vectors along with helper plasmids psPAX2 and pMD2.G (addgene Plasmid #12260 and #12259). Medium was collected 24 and 48 hours after transfection and filtered through a 0.45 µm filter. Virus was then added to 10^7 CHO-K1 (ATCC CCL-61, Manassas, VA) cells at a multiplicity of infection (MOI) of 0.01 or 0.02.

6.3.4 PCR-Based Integration Site Analysis

Integration site analysis for the lentivirus-based clones was performed using the Lenti-X Integration Site Analysis Kit (Takara Bio USA, Mountain View, CA). A diagram of this process is shown in Figure 6-2.

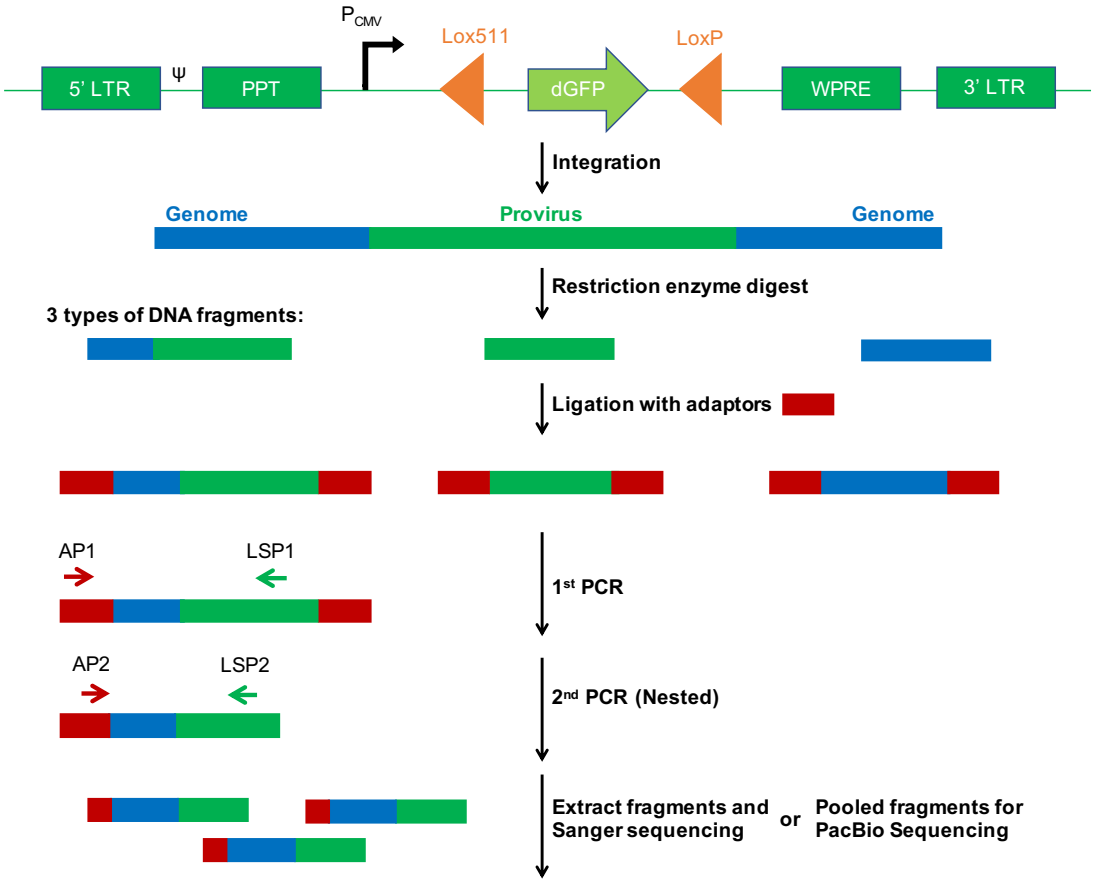

Figure 6-2. PCR-based integration site analysis. Genomic DNA is first fragmented using restriction enzymes and then ligated with an adaptor. Two rounds of PCR using a lentivirus specific primer (LSP) and an adaptor specific primer (AP) are used to enrich for fragments containing both lentivector and genome sequences. Extracted fragments are then sequenced using Sanger sequencing.

Briefly, genomic DNA (gDNA) was extracted from the cells using NucleoSpin Tissue Kit (Takara Bio USA, Mountain View, CA). The gDNA was then digested with three restriction enzymes: DraI, SspI, and HpaI. Digested DNA was further purified using NucleoSpin Gel and PCR Clean-Up Kit (Takara Bio USA, Mountain View, CA). After purification, GenomeWalker adapters were ligated to the fragmented gDNA. Two successive rounds of PCR were then performed to amplify gDNA fragments containing vector sequence using the Advantage 2 PCR Kit (Takara Bio USA, Mountain View, CA). First, PCR was done using the Lentiviral Specific Primer 1 (LSP1), which binds to the ψ packaging sequence on the inserted lentiviral vector, and Adapter Primer 1 (AP1), which binds to the ligated adapter. The PCR reaction was diluted, and then added to a second reaction using nested primers LSP2 and AP2, found downstream of the original set used

for amplification. Gel electrophoresis was used to visualize the nested PCR reaction, and in the case of single integration of the lentivirus, single bands were extracted from the gel and purified using NucleoSpin Gel and PCR Clean-Up Kit (Takara Bio USA, Mountain View, CA). The purified PCR product was Sanger sequenced at the University of Minnesota Genomics Center (Saint Paul, MN).

6.3.5 PacBio Sequencing

Sequencing of enriched PCR fragments was performed by the National Center For Genome Resources (NCGR, Santa Fe, NM). Library was prepared using the PacBio DNA 2kb Template Prep Kit. The sequencing was performed in one SMRT cell at NCGR. This gave approximately 770K reads per sample, with an average read length of 1.6 kbp. Further processing of PacBio data is detailed in Figure 6-3.

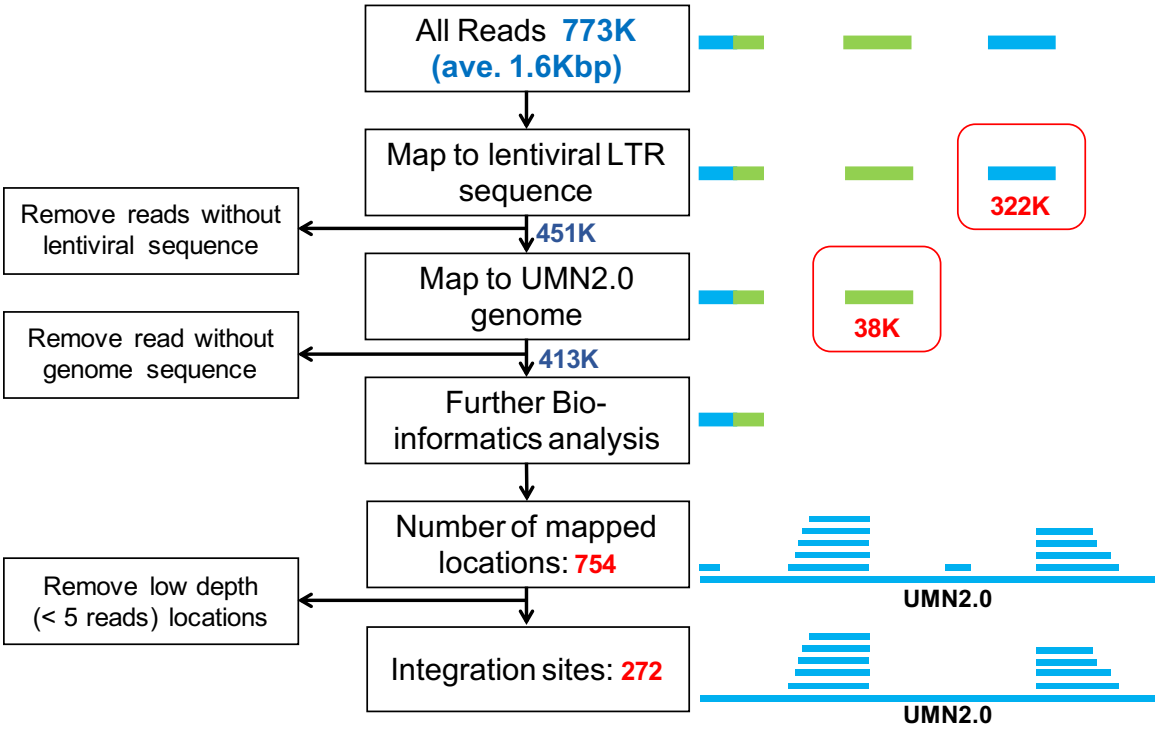

Figure 6-3. PacBio Sequencing bioinformatics analysis. Reads which did not contain both lentiviral and genome sequences were filtered out, and the remaining reads were mapped to the genome to determine the integration sites. Low confidence reads were removed. Numbers shown are for the top 1% GFP^+ pool; similar results were found for the bottom 10% pool.

6.3.6 RNA-seq

RNA from CHO-K1 or 2C10 cells was extracted using RNeasy Mini Kit (Qiagen, Valencia, CA). Strand-specific RNA-seq Illumina library creation and HiSeq 2500 High-output 125-bp paired-end Illumina sequencing was done at the University of Minnesota Genomics Center (Saint Paul, MN).

Sequencing yielded 21 million paired end reads. Data was trimmed to remove TruSeq3 adapter sequences using Trimmomatic. Reads were then mapped to the Chinese Hamster Genome (UMN 2.0) [85] using bowtie2 [150] under parameters --read-realign-edit-dist 0 --library-type=fr-firststrand --b2-very-sensitive. For 100 kbp windows, data was normalized by Reads per Kilobase per Million reads Mapped (RPKM). For analysis of gene expression, data was normalized by Fragments per Kilobase per Million reads Mapped (FPKM) using the NCBI *Cricetulus griseus* Annotation Release 101.

Gene expression in terms of FPKM was quantified using Cufflinks. FPKM for GAPDH in 2C10 was calculated separately due to the presence of a GAPDH pseudogene (LOC100757828) with high homology to GAPDH within the Chinese Hamster Genome. To better estimate the GAPDH FPKM, reads mapping to both this pseudogene region and the correct location were re-mapped to the single scaffold only containing the correct GAPDH gene (Scaffold 731). The FPKM of GAPDH was calculated by Cufflinks and scaled back to the original data set.

6.3.7 Assay for Transposase Accessible Chromatin using Sequencing (ATAC-seq)

$5x10^4$ CHO-K1 cells were trypsinized and nuclei were prepared (as detailed in [146]). Transposition was carried out using the Nextera DNA Library Preparation Kit (Illumina, San Diego, CA), and tagmented DNA was purified using Qiagen MinElute spin columns. The purified library was sequenced using the Illumina HiSeq 2500 High-Output platform at the University of Minnesota Genomics Center (Saint Paul, MN).

Sequencing yielded 150 million paired end reads (2x50 bp), which corresponds to a 6x coverage of the Chinese Hamster genome. The data was first trimmed to remove Nextera transposase adapters using Trimmomatic, and mapped to the Chinese Hamster mitochondrial genome using bowtie2 [150] under standard parameters to remove reads from mitochondrial DNA. The remaining reads were mapped to the Chinese Hamster

Genome (UMN 2.0) [85] using bowtie2 under parameters -N 1 -L 20 -i S,1,0.50 -D 20 -R 3 -I 38 -X 2000. Alignments with MAPQ < 10 were discarded using samtools and duplicates were removed using Picard. Reads aligning to the + strand were shifted by +4 bp and reads aligning to the – strand were shifted by -5 bp to reflect the true cut sites of the Tn5 transposase [145]. Additionally, reads were un-paired to observe open chromatin instead of nucleosome profiling. ATAC-seq peaks were called using the MACS2 algorithm [151] and the parameters -p 0.000000001 --nomodel --shift -25 --extsize 50. Read coverage was normalized to 100 million total mapped reads via RPKM.

6.3.8 Dual RMCE

Dual RMCE was performed in 24-well culture plates. Cells were seeded at $4x10^4$ cells/well and allowed to grow overnight. Cells were transfected with 0.2ng of donor vector and 0.4ng of the recombinase vector using DNA-In CHO (MTI-GlobalStem, Gaithersburg, MD). $mCherry^+$ and GFP^+ cells were sorted using FACS (FACSAria II, BD, Franklin Lakes, NJ) 5 days after transfection.

6.3.9 qRT-PCR for quantifying transcript levels

Transcript levels of the top 52 IgG producing clones from lentiviral infection were quantified using the Cells-to-CT 1-Step Power SYBR Green Kit (Life Technologies, Carlsbad, CA).

For other qRT-PCR experiments, RNA was extracted from cells using RNeasy Mini Kit (Qiagen, Valencia, CA), and cDNA synthesis was done using SuperScript III First-Strand cDNA Synthesis SuperMix for qRT-PCR (EPO and TNFR transgene replacement experiments) or SuperScript III First-Strand cDNA Synthesis System (IgG and GFP lentiviral clone experiments). qRT-PCR was performed using SYBR Select Master Mix (Applied Biosystems, Foster City, CA).

6.3.10 Enzyme-linked immunosorbent assay (ELISA)

IgG from cell supernatants was captured using Anti-Human IgG, Fc specific antibody (Sigma-Aldrich I3391, St. Louis, MO) coated onto Nunc MaxiSorp ELISA Plates (BioLegend 423501, San Diego, CA). Monoclonal Anti-Human IgG-Alkaline Phosphatase antibody (Sigma-Aldrich A2064, St. Louis, MO) was used for detection, and signal was

developed using Sigmafast p-Nitrophenyl Phosphate tablets. ImmunoPure Human Gamma Globulin (Invitrogen 31879, Carlsbad, CA) was used as the standard.

TNFR-Fc fusion protein in cell supernatants was captured using TNFRII Antibody (R&D Systems MAB726, Minneapolis, MN) coated onto Nunc MaxiSorp ELISA Plates (Biolegend 423501, San Diego, CA). Mouse Anti-Human IgG_1 Fc Fragment Specific Biotin Conjugate Antibody (Millipore/Calbiochem 411543, Burlington, MA) was used for detection, and signal was developed using Streptavidin-HRP (R&D Systems DY998, Minneapolis, MN) and Color Substrate Solution (R&D Systems DY999, Minneapolis, MN). TNFR-Fc purified from producing cell lines was used as a standard.

Erythropoietin in cell supernatants was quantified using the LEGEND MAX Human Erythropoietin (EPO) ELISA Kit (BioLegend, San Diego, CA).

6.3.11 Growth Characterization of Cells

Growth of dual RMCE derived cell clones was characterized by seeding a 24 well plate at 1 x 10^4 cells per well and collecting 3 wells each day for 7 days for cell counting. Media was changed on all wells on day 4. Supernatants on day 7 were collected for ELISA.

6.3.12 Statistical Analysis

The non-parametric Kolgomorov-Smirnov test [152] for testing if two samples are derived from the same parent distribution was used to test for significant differences in the distributions of RNA-seq activity between the integration sites from different GFP^+ populations. Multiple peaks could be seen in these distributions. The Kolgomorov-Smirnov test allows for distinguishing differences in the shape of the distributions, in addition to differences in the median values. As ATAC-seq RPKM distributions were non-normal, but more unimodal, differences in the median ATAC-seq activity between these populations were quantified using the non-parametric Mann-Whitney Test. Both the Kolgomorov-Smirnov and Mann-Whitney tests gave similar significance values for the comparisons examined. Significant differences in qRT-PCR transcript levels were determined using a Student's T-test. All statistical tests utilized a significance threshold of $\alpha = 0.05$, and analysis was performed using the stats package in R.

6.4 Results

6.4.1 Lack of correlation between transgene expression and genetic/intergenic status of integration sites

To establish a system for characterizing transgene integration sites, CHO-K1 cells were infected with a lentiviral vector containing destabilized GFP (dGFP) at a low MOI of 0.01 to reduce multiple integration events (Figure 6-1A). We used dGFP with a half-life of 2 hours as a reporter [153] to better reflect transcriptional activity. Forty-eight hours after infection, the top 1% and bottom 10% of GFP^+ cells were sorted using FACS and subsequently cultured for 4 weeks to expand the cell population for further characterization. The integration site of the lentivirus genome was PCR amplified using primer pairs flanking the junction of the CHO genome and the lentivector, as described in the Experimental Section. Amplicons from the two populations were isolated for PacBio sequencing (Figure 6-2). Sequencing resulted in approximately 770k reads per sample, with an average read length of 1.6 kbp. Reads containing both genome and vector regions were isolated using BLASTn. Initial filtering removed reads with less than a 400 bp alignment to the lentiviral LTR region. The remaining reads were aligned to the Chinese hamster genome, and the integration location for each read was identified (Figure 6-3).

For the top 1% population, 41.7% and 4.9% of reads mapped exclusively to lentiviral sequences and the host cell genome respectively. From the remaining 53.4% of reads, 272 integration sites on host genome were identified after low confidence regions were removed. Over half (64.3%) of the integration sites were located in intergenic regions, 34.9% were within introns, and the remaining 0.7% of sites within exons (Figure 6-4A). In the bottom 10% population, 254 integration sites were identified. These integration sites were distributed over intergenic, intron, and exon regions in a similar fashion as those in the top 1% population. As a measure of the abundance of these features within the genome, ten thousand randomly selected sites within the genome were examined, and the distribution of gene features was similar to the top 1% and bottom 10% pools. With respect to the gene features of the integration location, the top 1% and bottom 10% thus did not appear to be different from each other or the distribution of these features in the genome.

6.4.2 Isolation of integration sites for high expressing GFP clones

In a second experiment, we infected CHO-K1 cells with the same dGFP lentivirus at an MOI of 0.01-0.02, and cells in the top 1% of green fluorescence were single-cell cloned and expanded. A total of 89 high-GFP intensity clones were obtained. The junction region of the integration site was amplified by PCR, and the amplified DNA was separated by agarose gel electrophoresis to identify the number of integration sites, or the number of copies of the dGFP lentivirus in each clone. We identified 34 clones as having a single lentivirus integrated into the CHO genome. The rest of the clones were similarly distributed between having 2, 3, or 4+ copies. Gel bands from the single copy clones were excised for DNA sequencing using the Sanger method, and the integration sites were then identified by BLASTn against the Chinese hamster genome. The integration sites for these single copy clones were mostly in introns (79.4%), and the remaining 20.6% were in intergenic regions (Figure 6-4A). This distribution of gene features is different from integration locations of the top 1% pool, bottom 10% pool, and random locations, which were mostly in intergenic regions. It should be noted that the GFP expression in all clones is attributable to a single integration site while some cells in the pool contain multiple copies. This may account for the difference observed.

6.4.3 Activity of Region Surrounding Integration Site Correlates with GFP Expression

We next asked the question whether high transcriptional activity and genome accessibility in the region surrounding the locus of integration coincides with high transgene expression as reflected in GFP intensity. As an assessment of the regional transcriptional activity and accessibility, RNA-seq and ATAC-seq data from CHO-K1 cells was used to characterize the region 50 kbp upstream and downstream of the integration site. The sum of all RNA-seq and ATAC-seq reads were normalized to Reads per Kilobase per Million Reads (RPKM) and tabulated.

Density distributions of the regional transcriptional activity of each integration site for the three populations of GFP^+ cells (top 1% single-copy clones, top 1% pool, and bottom 10% pool) were plotted in Figure 6-4B. As a comparison, 10,000 sites randomly selected from the genome were also plotted. The distributions of the RNA-seq data were

not unimodal, as there were many integration sites with low RPKM in addition to a broad peak around Log_2[RPKM] of zero. Hence, a Kolmogorov-Smirnov test for the equality of two different distributions was used to compare the different data sets (Table 6-1). In general, the top 1% pool and top 1% clone distributions were shifted towards higher activity, while the bottom 10% pool and random population were skewed towards lower activity. The transcriptional activity of the integration sites for the bottom 10% population had a similar distribution to the random population (p = 0.093), while that of both the top 1% pool and top 1% clones were significantly different from the bottom 10% as well as the random population (p <0.001 for all comparisons).

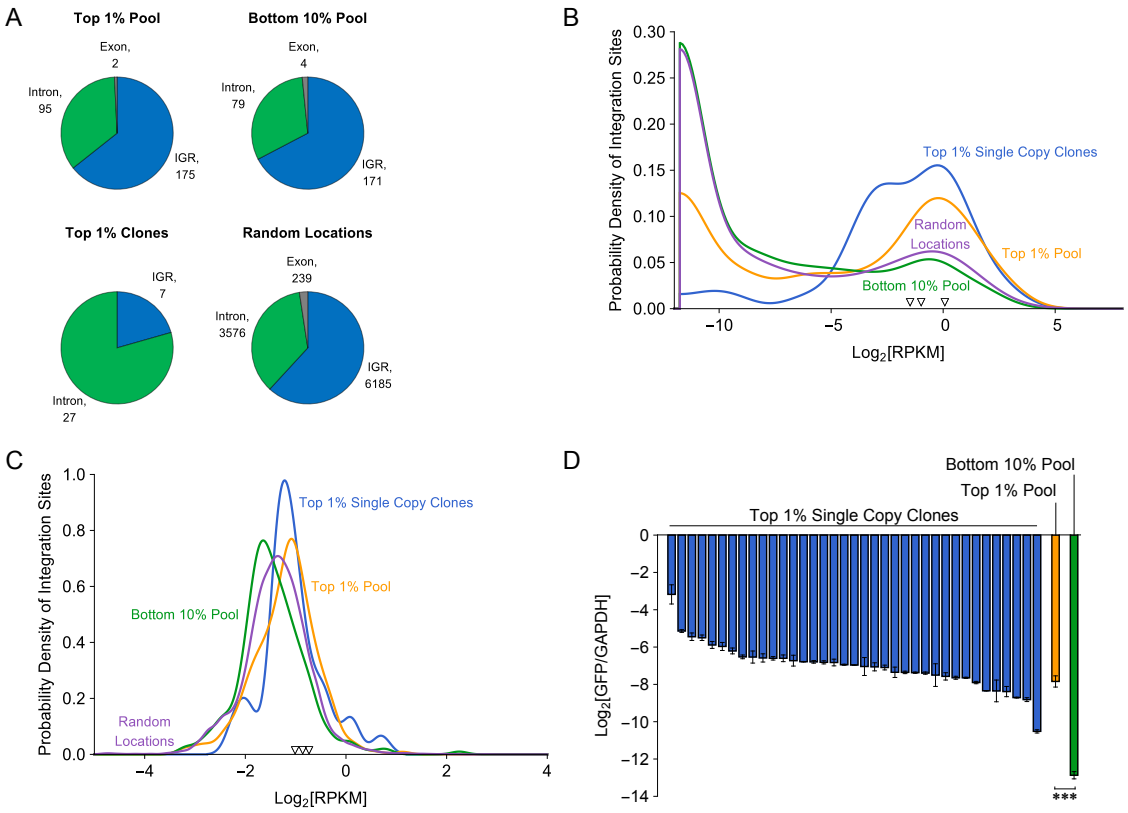

Figure 6-4. Transgene integration into active regions with open chromatin. (A) Gene features at integration sites found for the different populations of GFP^+ cells (top 1% pool, bottom 10% pool, and top 1% single copy clones) and random locations. (B) & (C) Probability density of log_2[RPKM] for RNA-Seq (B) or ATAC-Seq (C) data across a 100 kbp window centered on integration sites found for different cell populations. (Blue: top 1% single copy clones, N = 34 integration sites; Orange: top 1% pool, N = 272 integration sites; Green: bottom 10% pool, N = 254 integration sites; Purple: 10,000 random locations). Points (▽) on the baseline denote Log_2 [RPKM] values for three single copy IgG clones. (D) qRT-PCR data for transcript level of GFP among top 1% single copy GFP producing clones, top 1% GFP pool, and bottom 10% GFP pool, as measured by log_2 fold change relative to GAPDH (***: $p < 0.001$).

The population density distribution of the top 1% single copy clones had a similar distribution to the top 1% pool, except skewing towards higher expression (Figure 6-4B). For cells in the top 1% pool, a fraction of the population is likely to have more than one copy. Thus, the GFP intensity may have resulted from combined expression from multiple loci, or from a major, high-expressing locus with accompanying inactive loci.

Table 6-1. Kolmogorov-Smirnov test for comparison between RNA-seq RPKM distributions and Mann Whitney Test for the comparison of ATAC-Seq RPKM distributions in the 100 kbp window surrounding the integration sites. $p < 0.05$ denotes a significant difference between the distributions. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$)

RNA-seq RPKM distribution comparison	**K-S test *p*-value**
Top 1% pool > bottom 10% pool***	1.90×10^{-12}
Top 1% clones > bottom 10% pool***	7.03×10^{-11}
Top 1% clones ≠ top 1% pool**	4.13×10^{-03}
Top 1% pool > random locations***	2.42×10^{-18}
Bottom 10% pool ≠ random locations	0.093
Top 1% clones > random locations***	3.65×10^{-10}
ATAC-seq RPKM distribution comparison	**Mann-Whitney test *p*-value**
Top 1% pool > bottom 10% pool***	2.57×10^{-08}
Top 1% clones > bottom 10% pool***	1.76×10^{-04}
Top 1% clones ≠ top 1% pool	0.509
Top 1% pool > random locations***	8.88×10^{-08}
Bottom 10% pool ≠ random locations**	7.75×10^{-03}
Top 1% clones > random locations**	2.38×10^{-03}

For integration sites located within genes, we also examined whether there was a relationship between GFP fluorescence and the RNA-seq activity of those genes. In this case, the distributions for the top 1% and the bottom 10% pools were not significantly different (Figure 6-5 and Table 6-2, $p = 0.129$). The data thus implies that cells isolated via sorting for high GFP fluorescence were more likely to have integrated into general regions of high transcriptional activity, while the expression of the specific gene at the integration site may not be relevant.

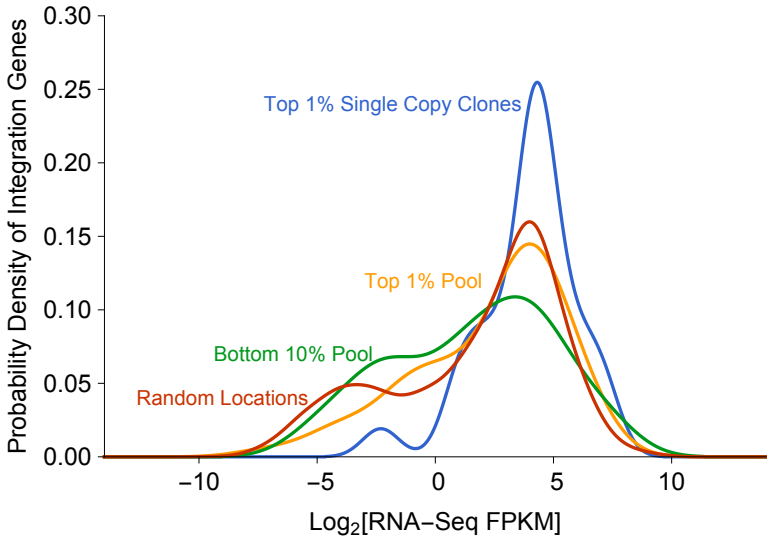

Figure 6-5. Probability density of integration sites within genes based on gene RNA-Seq FPKM. Integration sites outside of gene regions were excluded from this analysis. (Blue: top 1% single copy clones, n = 27 integration sites; Orange: top 1% pool, n = 70 integration sites; Green: bottom 10% pool, n = 53 integration sites; Red: 10,000 random locations).

Table 6-2. Kolmogorov-Smirnov test for comparison between gene RNA-seq FPKM distributions for integration sites within genes. $p < 0.05$ denotes a significant difference between the distributions. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$)

RNA-seq Gene FPKM Distribution Comparison	K-S test *p*-value
Top 1% Pool > Bottom 10% Pool	0.129
Top 1% Clones > Bottom 10% Pool*	0.005
Top 1% Clones ≠ Top 1% Pool*	0.037
Top 1% Pool > Random Locations	0.252
Bottom 10% Pool ≠ Random Locations	0.602
Top 1% Clones > Random Locations**	0.004

ATAC-seq data was quantified as described in the Experimental Section for each integration site identified. The number of ATAC-seq reads from +50 kbp to –50 kbp of each integration site were normalized and reported as ATAC-seq RPKM. The population density distribution for each of the four data sets, expressed as $\log_2$ transformed ATAC-seq RPKM, is shown in Figure 6-4C. Significance of differences among the median of these distributions was examined using the non-parametric Mann-Whitney Test (Table

6-1). All three GFP^+ populations were significantly different from the random set ($p < 0.005$). Additionally, the distribution of both top 1% sets, the single copy clones as well as the pool, had significantly higher ATAC-seq expression over the bottom 10% pool ($p < 0.001$). No significant difference was observed between the single copy and pooled top 1% cell integration sites ($p = 0.51$).

qRT-PCR was then used to measure the GFP transcript level of the three GFP^+ populations (Figure 6-4D). The pooled top 1% of GFP^+ cells had significantly higher transcript of GFP compared to the pooled bottom 10% (2^5 (32) fold higher, T-test, $p < 0.001$). Among the 34 clones isolated from the top 1% of GFP^+ cells, the expression of GFP falls within a narrow range. Their transcription level of GFP was similar to the top 1% pool, which is expected as the clones were isolated from a similar population of top 1% GFP positive cells. The data thus suggests a positive correlation between GFP fluorescence intensity, GFP transcript level, and integration into regions with higher transcriptional activity and accessibility as measured by RNA and ATAC sequencing.

6.4.4 Single Copy IgG Producing Cell Line

Having shown that the top GFP^+ cells obtained by infection with a low MOI of lentivirus would have a higher probability of landing the transgene into a transcriptionally active and accessible region of the CHO genome, we set out to use the same method to integrate a lentiviral vector containing IgG heavy chain and light chain genes. It has been previously shown that GFP fluorescence correlated with titer of a co-expressed product gene [137], which further supports our method. The heavy chain and light chain genes were followed by an internal ribosome entry site (IRES) to drive translation of dGFP (Figure 6-1B). The vector also contained recombination sites for flippase and cre recombinase flanking the IgG sequence, such that the product gene could be replaced at a later date.

CHO-K1 cells were infected with IgG-IRES-dGFP lentivirus at an MOI of 0.02. Forty-eight hours after infection, the top 10% of GFP^+ cells were collected using FACS (Figure 6-6), cultivated for 10 days, and then sorted again. The top 0.1% of GFP positive cells (2,982 cells) were collected into separate wells via limiting dilution for further cell expansion. After 14 days, supernatant was collected from the 303 surviving clones and the IgG level was quantified using ELISA (Figure 6-7).

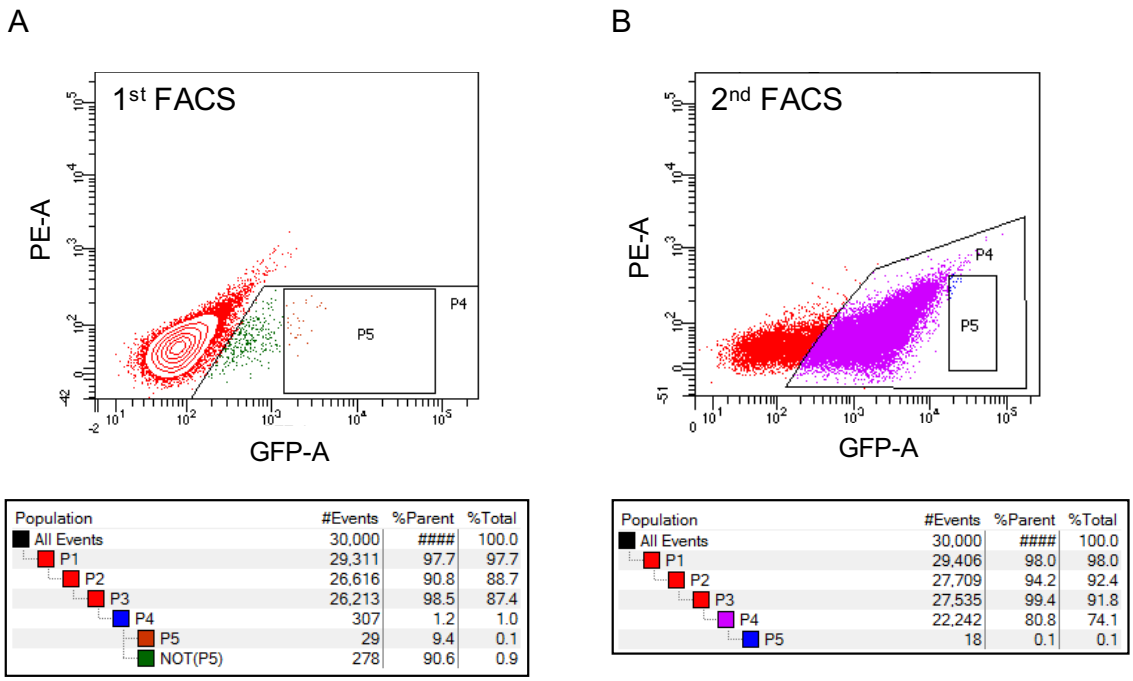

Figure 6-6. FACS to isolate top GFP producing cells 48 hours after lentiviral infection. Plots of PE vs. GFP fluorescence. (A) 1st FACS to isolate top ~10% of GFP$^+$ cells. (B) 2nd FACS to isolate top 0.1% of GFP$^+$ cells.

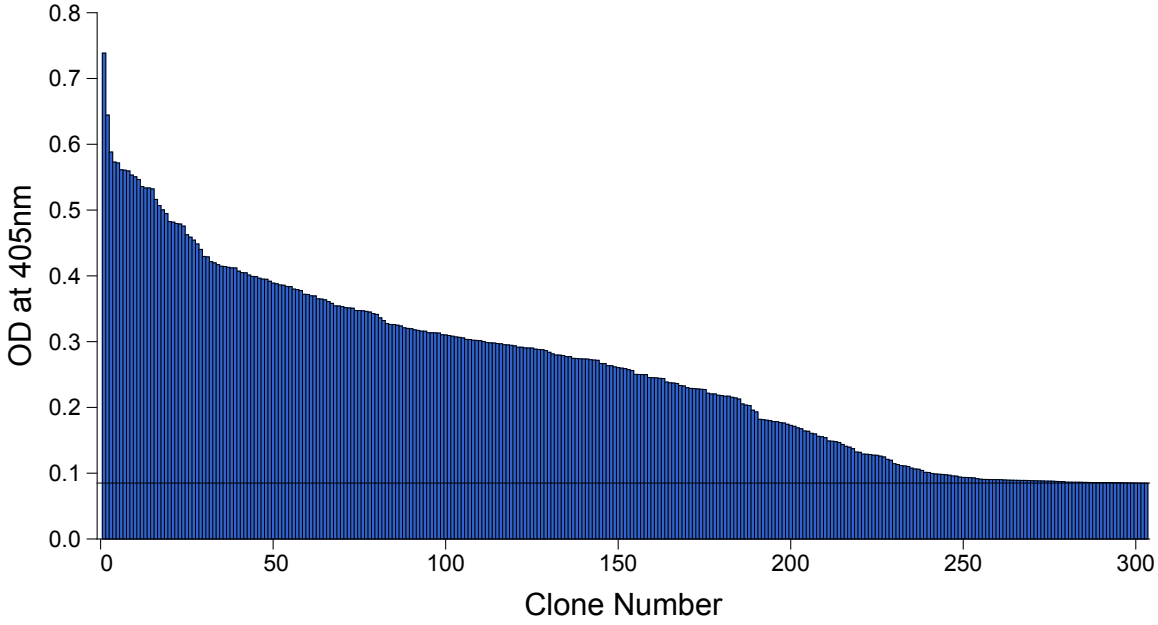

Figure 6-7. ELISA data for 303 single cell clones isolated from top 1% of GFP$^+$ cells after IgG-IRES-dGFP lentiviral infection. Black line denotes OD at 405nm measured for blank control.

The 52 clones with the highest IgG levels were further expanded in 6-well plates for 5 days before the IgG titer and transcript levels were quantified. The IgG titer in the supernatant had a large range (Figure 6-8A). Transcript levels for heavy chain and light chain of IgG were quantified using a one-step qRT-PCR method that, which although consistently 1-2 cycles different from the standard method we used, was better suited for screening a large number of clones, as described in the Experimental Section (Figure 6-8B). All heavy and light chain transcript levels were similar for most clones. The expression level ranged from $2^{0.7}$ to $2^{5.7}$ (1.62 to ~52) fold higher than GAPDH for light chain and from $2^{0.5}$ to $2^{6.8}$ (1.4 to ~111) fold higher for heavy chain among different clones.

The number of copies of lentivector integrated in each clone was then determined using the PCR-based integration site analysis method followed by gel electrophoresis. Out of the 52 top IgG producing clones, 8 (15.4%) had a single copy, 17 (32.7%) had two copies, 10 (19.2%) had three copies, and 17 (32.7%) had more than 3 copies (Figure 6-8A and Figure 6-9A). Notably, copy number did not seem to be directly correlated with IgG titer. Some single copy clones had the same titer or even a higher titer than clones with greater than three copies.

6.4.5 Three high IgG secreting clones express high level of transcripts

Three of the single copy clones, 2C10, 3A10, and 4B1, were characterized further. Both light chain and heavy chain transcripts were at levels higher than GAPDH in all three clones (Figure 6-9B). Specific productivity of IgG for these clones ranged from 18 – 29 pg/cell/day (Figure 6-9C).

The integration site of the vector in all three clones was identified. ATAC and RNA-seq of the host cell, CHO-K1, were used to examine the accessibility and the transcriptional activity in the region of these three integration sites (Figure 6-9D-F). In the 100 kbp region surrounding the integration site, all three clones had high RNA-seq and ATAC-seq read density (Figure 6-4B and C, marked ▽). This implies the clones integrated into active regions with open chromatin, which agrees with our findings from the original GFP lentivirus experiment showing high expression from active sites.

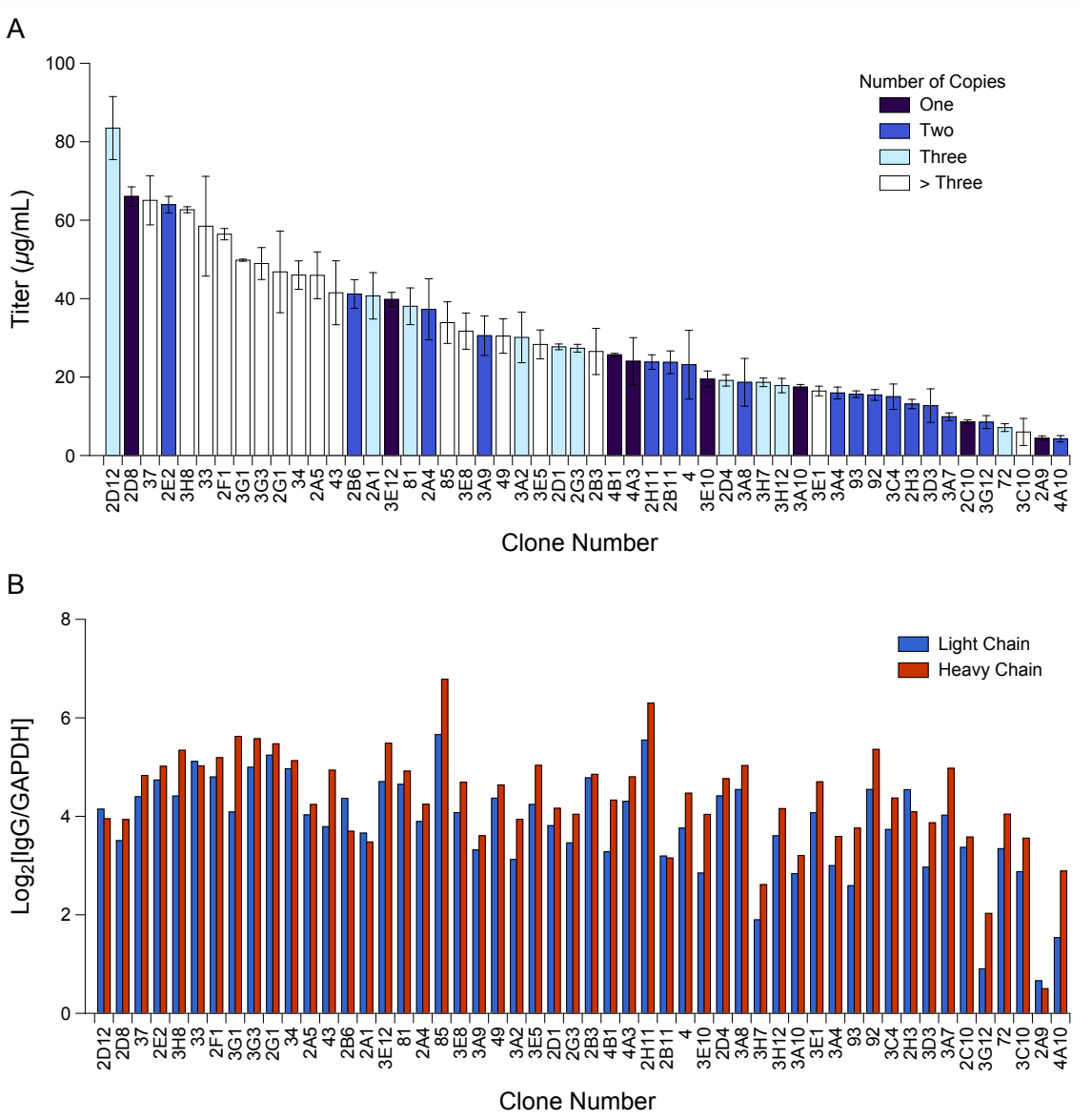

Figure 6-8. Single Copy IgG Cell Lines. (A) IgG titer from of top 52 IgG producing clones as measured by ELISA. Supernatant was collected from adherent culture of cells in 6 well plates after five days. Clones were selected based on ELISA data at the 96-well stage of single cell cloning. Color of bar corresponds to number of copies of lentivector integrated as determined by integration site analysis. (B) Transcript level of IgG heavy chain and light chain for top 52 clones measured using 1-step qRT-PCR. Results reported in terms of $\log_2$ fold change relative to GAPDH.

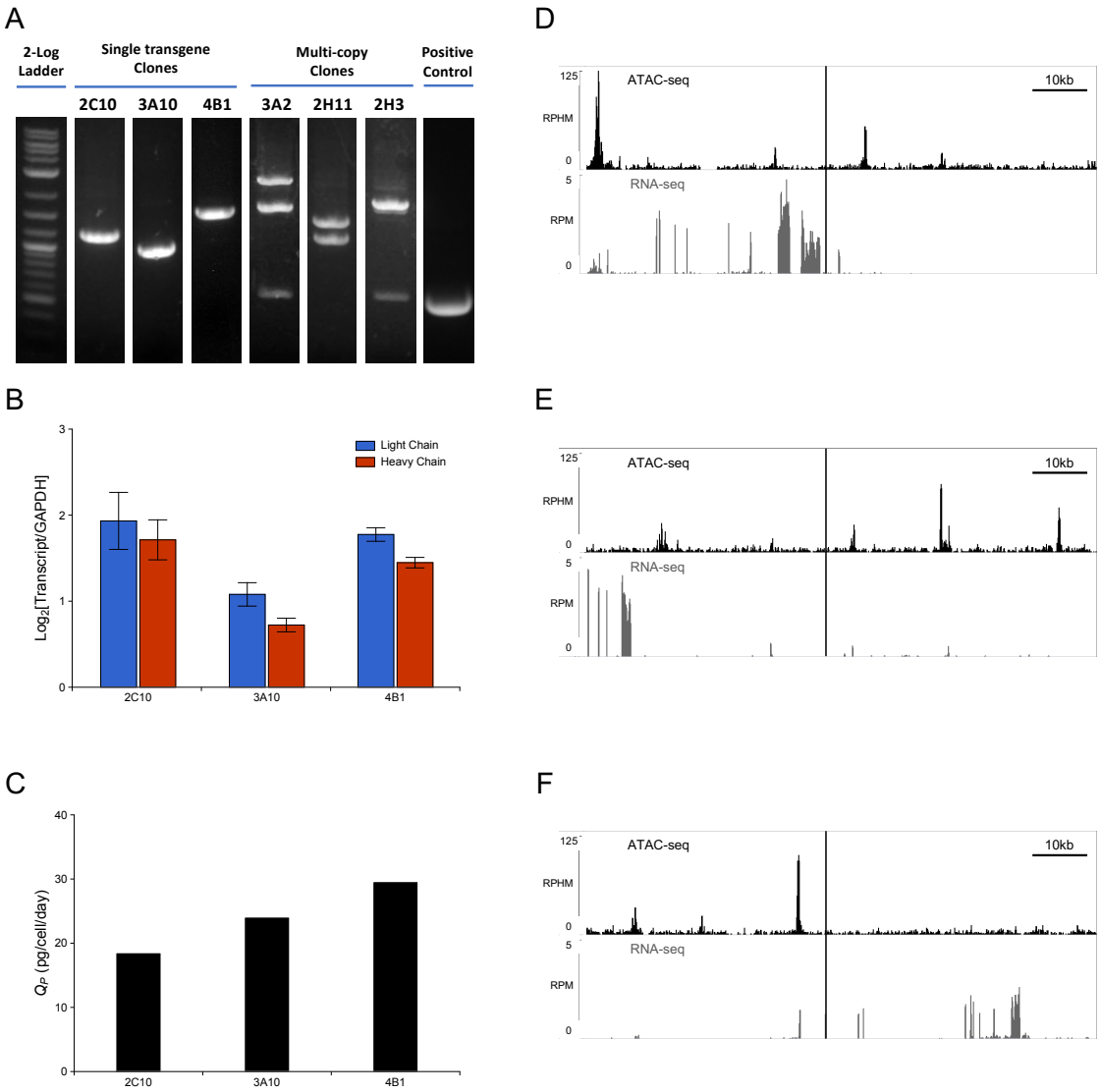

Figure 6-9. Characterization of single copy IgG producing cell lines. (A) Gel-electrophoresis of secondary PCR products from integration site analysis. Number of bands corresponds to number of integration sites. (B) qRT-PCR data measuring IgG light chain (LC) and IgG heavy chain (HC) transcript level in three single copy clones with different integration sites. Measured in terms of log2 fold change relative to GAPDH. (C) Specific productivity of single copy clones assayed in adherent culture as measured by ELISA. Supernatant was collected on eighth day of culture, two days after media replacement. N = 2 wells per cell line. (D-F) ATAC-Seq and RNA-seq pile up from CHO-K1 for 100 kbp window surrounding integration sites of three single copy clones (D: 2C10, E: 3A10, F: 4B1) as visualized in IGV. ATAC-seq data is normalized to Reads Per Hundred Million (RPHM) and RNA-seq data is normalized to Reads Per Million (RPM). Vertical black bar denotes integration site.

6.4.6 Transgene Replacement Using Dual Recombinase Mediated Cassette Exchange (RMCE)

The 2C10 cell line producing IgG was used to generate cell lines producing a new product using dual recombinase mediated cassette exchange (RMCE). For recombination, we employed a recombinase vector encoding flippase and cre recombinase. 2C10 was transfected with a donor vector containing either the sequence for Darbepoetin alfa, an engineered variant of Erythropoietin (EPO), or the sequence for Entanercept, a TNFR-Fc fusion protein, along with the recombinase vector. The donor vector contained puromycin resistance and mCherry, and the sequence for insertion was flanked by Lox511 and minimal FRT recombination sites that match those on the original lentivector. This vector was designed without a promoter for the mCherry gene, such that random integrants would not be selected during sorting (Figure 6-10A). The product gene was also promoter-less, meaning that expression of the product gene from random integration was highly unlikely.

$mCherry^+$ and GFP^+ cells were isolated using FACS five days after transfection. (Figure 6-11). Clones were obtained by single cell cloning, and correct recombination was confirmed with PCR across both the lox511 and minimal FRT junctions (Figure 6-12). Five EPO producing clones were obtained, and production of the new protein was verified by ELISA. Low cloning efficiency after sorting led to the isolation of only one TNFR-Fc producing clone.

The six clones exhibited very similar growth rates, although they reached somewhat different cell densities (Figure 6-10B). Transcript levels of EPO varied among the different clones (Figure 6-10C). The transcript levels of EPO and TNFR were lower than GAPDH, ranging from $2^{0.9}$ to $2^{4.4}$ (1.9 to ~21) fold lower than GAPDH for EPO, and $2^{1.3}$ (2.5) fold lower than GAPDH for TNFR. Specific productivity of EPO varied and followed the same trend as the transcript data (Figure 6-10D). The TNFR-Fc producing clone had slightly higher specific productivity than the EPO clones.

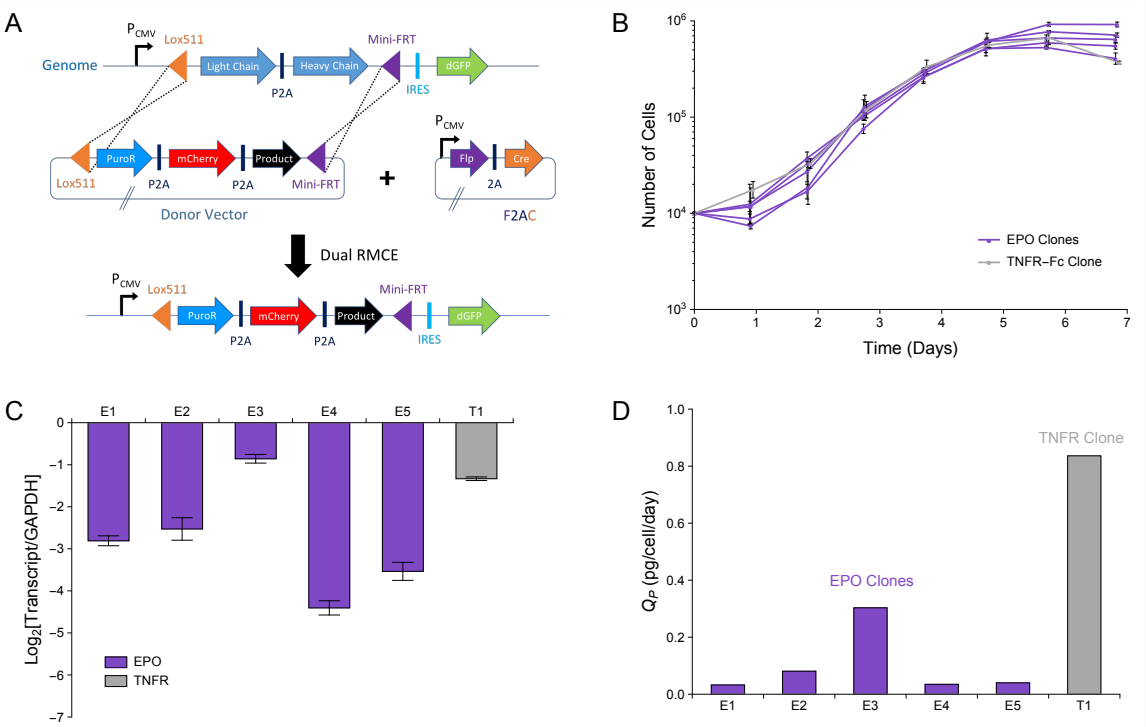

Figure 6-10. Replacement of the IgG Product Gene using Dual RMCE. (A) Schematic of dual RMCE process. Donor vector is promoter-less and contains Lox511 and minimal-FRT sites that match those on the genome. F2AC is the recombinase expression vector containing flippase and cre recombinase. (B) Growth curve for EPO and TNFR producing clones obtained from dual RMCE. Cells were grown in adherent culture for seven days, and media was replaced on day four. (C) qRT-PCR data measuring transcript level of EPO and TNFR, as denoted by $\log_2$ fold change relative to GAPDH. (D) EPO and TNFR specific productivity as measured by ELISA. Supernatant was collected on day seven of the growth curve from (B), three days after media replacement. N = 3 wells per cell line.

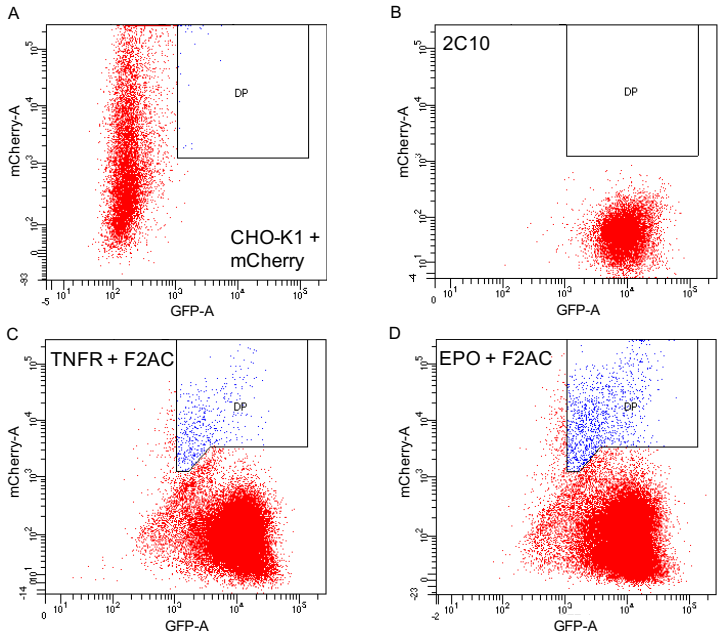

Figure 6-11. FACS to isolate mCherry$^+$ and GFP$^+$ cells after dual RMCE, 5 days after transfection. Plots of GFP vs. mCherry fluorescence. (A) CHO-K1 transfected with mCherry control plasmid. (B) Untransfected 2C10 cells. (C) 2C10 transfected with TNFR-Fc donor vector and F2AC recombinase plasmid. Approximately 1.1% of cells were mCherry+ and GFP+. After sorting and single cell cloning, the low cloning efficiency gave rise to only one TNFR-Fc clone. (D) 2C10 transfected with EPO donor vector and F2AC recombinase plasmid. Approximately 2.1% of cells were mCherry$^+$ and GFP$^+$. Five EPO producing clones were isolated after sorting and single cell cloning.

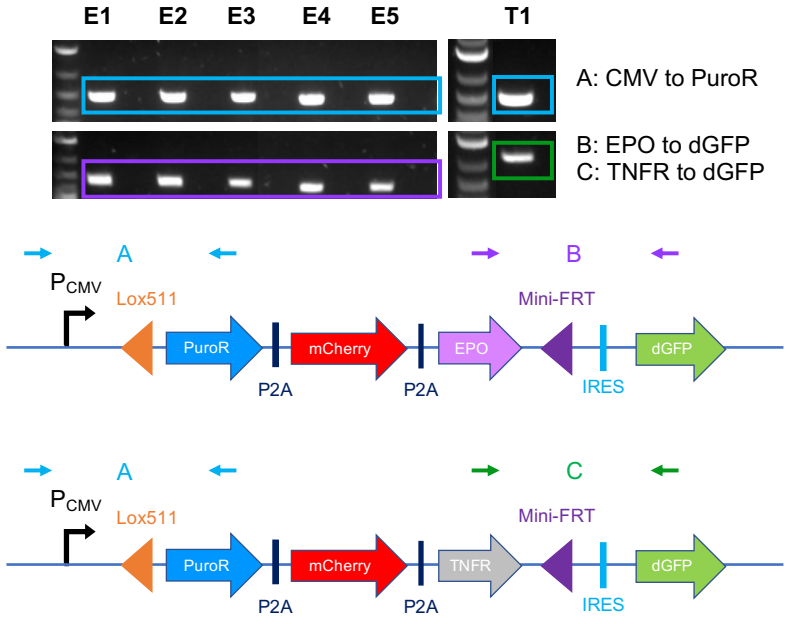

Figure 6-12. gDNA PCR to show correct integration of EPO and TNFR for dual RMCE derived clones. PCR product A confirms integration across the Lox511 junction, and PCR products B and C confirm integration across the Mini-FRT junction for EPO and TNFR swapped clones respectively.

6.5 Discussion

In this work, we used a lentiviral vector to generate cells with a very low number of transgenes integrated and showed that integration into regions with increased transcriptionally activity and accessible chromatin structure have a propensity to express the transgene at higher levels. We first demonstrated this using a dGFP reporter gene, and later showed that with an IgG transgene, the isolated high dGFP expressing clones also had higher IgG productivity. We employed a lentiviral vector to avoid the heterogeneous integration events associated with random integration of a transgene. In random integration, the vector is often truncated or ligated head-to-head or head-to-tail. Even if the copy number of the integrated transgene is determined, whether or not the gene is active or non-functional cannot be known without resorting to detailed genome sequencing. Whereas using a lentiviral vector, the integrated vector is largely intact. We also employed a dGFP that is tagged with a degradation domain to give a shorter half-life, which increases the dynamic range of its reporting of transcriptional activity [153]. An earlier effort in our lab using wild type GFP did not result in the isolation of cells with very high transgene transcript levels.

We noted that despite the very low MOI used, a large fraction of cells still had multiple copies of the transgene integrated. This could be attributed to aggregation of virus particles, resulting in multiple viruses entering a cell during a single infection event. Hence, the high IgG expressing clones isolated using the dGFP lentivirus were still examined for a single genome integration event via identification of the integration site. Using this strategy, all of the single copy transgene producing cell lines isolated had IgG transcript levels (both heavy chain and light chain) higher than GAPDH (Figure 6-8B), our reference standard for a highly expressed gene.

In examining the relationship between high transcriptional activity of the transgene (as indicated by GFP intensity) and the transgene integration site, we employed PacBio sequencing technology to obtain the integration site sequence in a cell pool. The 1.5 kbp reads obtained using PacBio allowed us to identify the vector-genome junction with reasonable confidence, which would not have been possible using methods that give only 100-200 bp reads. Additionally, the depth of sequencing allowed us to identify the genomic

integration site despite that nearly half of reads were non-specific (did not cover the vector-genome junction) and each integration site is present at a low level.

Many studies have looked into the mechanisms behind transcriptional activation. It has been found that active genes tend to be within regions that are sensitive to nucleases, such as DNAse I, and this sensitivity can extend many kilobases on either end of the gene [154]. Enhancers are known to recruit transcription factors and remodel regions of chromatin for increased transcriptional activity [155], and the majority of enhancers are found within 20-50 kbp of their target gene [113]. More recently, it has been shown that genes associated with large enhancer regions, called super-enhancers, are expressed at even higher levels [156]. Though we used ATAC-seq to evaluate pre-integrated sites within the genome, we can further use this chromatin accessibility data to look for enhancer or super enhancer regions in CHO to increase transgene expression in the future.

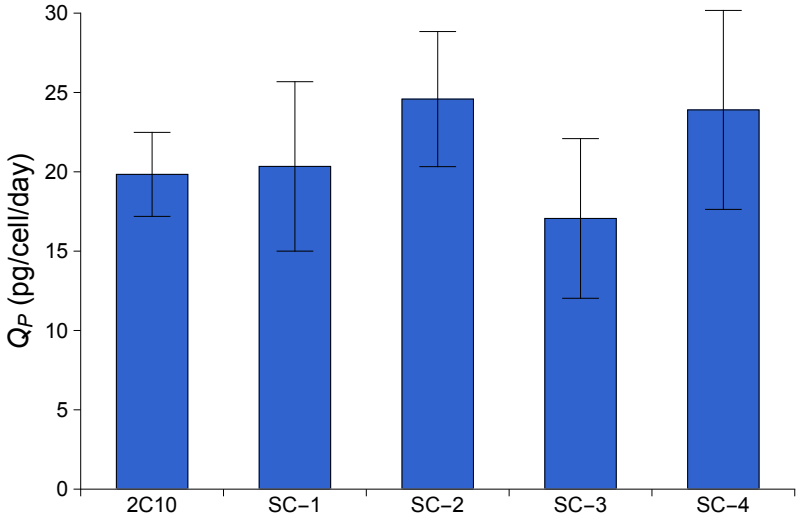

Figure 6-13. 2C10 and subclone (SC) specific productivity from ELISA. Data from cells grown in a 6 well adherent culture, n = 3 wells per cell line.

A very large fraction of the high GFP expressing and high IgG producing cells have transgenes integrated in regions with open chromatin. Heterochromatin invasion is a known mechanism of transcriptional silencing [157], which has led to the development of insulating elements to prevent the interference of regional chromatin upon gene integration [158]. In production cell lines, heterochromatin silencing has been shown to be correlated with instability of recombinant gene expression [159]. Integration of a transgene into a locus remote from a heterochromatin region may reduce the probability of heterochromatin

invasion causing a reduction of the transcriptional activity of the transgene, thus enhancing the stability of the cell line. Four subclones of one of the single copy IgG producing cell lines, 2C10, were isolated and shown to have a similar specific productivity (Figure 6-13). These cells were obtained after subculturing for a long duration and their specific IgG productivity remained unchanged after approximately 35-55 population doublings, suggesting that the transgene expression was indeed stable.

However, when the IgG genes were replaced with EPO by RMCE in the same cell line, the transcript levels varied about ten-fold among the EPO producing clones, as did the specific productivity. The parental cell line, 2C10, was selected for its capability to produce IgG, and the characteristics needed to produce EPO may be different. This may contribute to the variability. Other studies using RMCE to insert a product gene into a defined landing pad have also seen variations among producing clones [19, 20, 160, 161]. Few had quantified the transgene transcript level, and each had utilized different products and integration sites. Nevertheless, the spread of the specific productivity distribution among swapped clones varied between different reports, ranging from narrow to quite wide. Single cell cloning may have contributed to the variability, which is a well-known phenomenon in CHO cells [162, 163]. From RNA-seq data of 2C10 and the relative expression of the product genes by qRT-PCR, we estimated that the transcript levels of GAPDH, IgG-H, and IgG-L genes are at the abundance level of the top 0.41%, 0.013% and 0.013% of expressed genes respectively (Figure 6-14). In comparison, the TNFR and EPO producing clones would have their product gene expressed at the level of the top 0.88% to 14.8% of genes. Even though there is some variability in their expression level, overall, the product transcript was still abundant. Therefore, RMCE was still effective in generating cell lines with a high transcript level, though the robustness of the system still requires further study. Whether those clones will also be stable, as the transgenes were integrated in an accessible region of the genome, is yet to be examined.

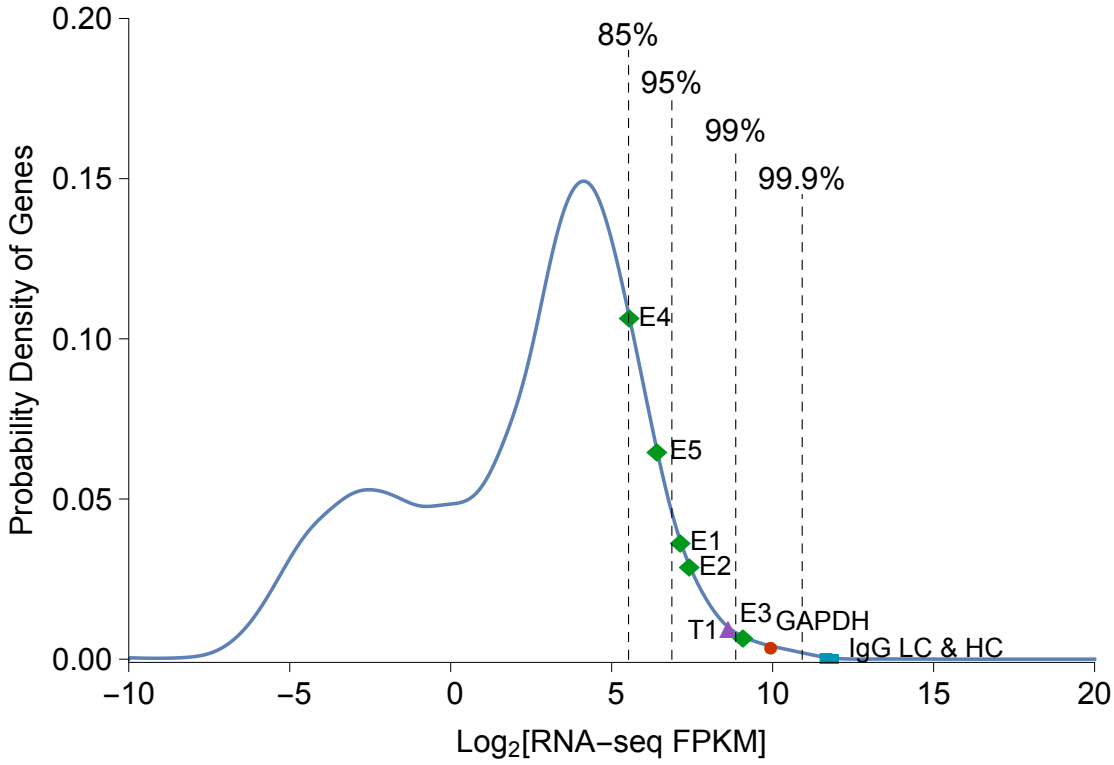

Figure 6-14. Probability density distribution of all expressed genes in cell line 2C10. Data reported in terms of Log_2 RNA-seq FPKM. Using the FPKM of GAPDH, and the relative expression of IgG and swapped product genes to GAPDH as measure by qRT-PCR, the expression of these genes in the different cell lines is plotted on the distribution curve. The 85th, 95th, 99th, and 99.9th percentiles of gene expression are marked on the graph with dotted lines. (Red ●: GAPDH expression in 2C10 from RNA-seq; Teal ■: IgG LC and HC relative expression in 2C10 from qPCR; Green ◆: EPO relative expression in EPO swapped clones E1, E2, E3, E4, & E5 from qPCR; Purple ▲: TNFR relative expression in TNFR swapped clone T1 from qPCR)

This study provides evidence that genome regions with high transcriptional activities and open chromatin structure are more amenable to high level expression of integrated transgenes. They may also provide the advantage of long-term stability in preserving transgene expression. With advances in genome engineering, one may insert a product transgene into a "favorable" genome region in many ways. We used a destabilized GFP vector and isolated high GFP$^+$ and product producing clones. By creating a landing pad (such as a LoxP site) or inserting RMCE sites into the transgene vector, one engineered clone could be used to create cell lines producing different products, as we demonstrated with the EPO and TNFR-Fc producing lines. Through RNA-seq and ATAC-seq, one may also identify regions of high preference and employ methods like CRISPR/Cas9 for

targeted integration of the product gene into the genome. In a separate study, we examined genome structural variants of CHO cells and uncovered regions that incur frequent structural changes [25]. Integration of a transgene into such regions may expose the production cell line to heightened risk of losing its productivity. By targeted integration such regions may be avoided. It is important to note that the integration site of the transgene largely influences only the transcriptional activity and stability. To become a high producing cell, a cell line must be endowed with many additional hyperproductivity traits, including high secretory capacity, enhanced redox balance, and improved energy metabolism [164]. Furthermore, it must have better growth kinetics under manufacturing conditions. Other factors, such as ribosomal occupancy, have also been shown to affect productivity [165]. Thus, by combining the method we presented with further cell line optimization, the process for obtaining a stable, hyperproducing cell line can be streamlined. Once such a cell line is established, the product gene can be exchanged for a new product using site-specific recombinases in order to retain both the integration locus and desirable hyperproductivity traits.

7 A combined modeling and cell engineering approach to reduce waste metabolite production in CHO cells

7.1 Introduction

Metabolism can have wide reaching effects on cells in culture. This is especially apparent in the growth of CHO cells, the primary cell line used to produce biologics. Changes to metabolic behavior during a culture can reduce the productivity of the culture [166] and affect the resulting product quality [36]. Additionally, waste metabolites such as lactate tend to be growth inhibitory, causing a cascade of undesirable effects on cell behavior and productivity.

Metabolically engineering the cells to overcome their natural limitations has long been of interest to the field of cell culture bioprocessing. Different pathways and enzymes have been targeted, including the apoptosis pathway [167], various enzymes in the TCA cycle [168], pyruvate dehydrogenase kinase (PDK) [169, 170], and lactate dehydrogenase (LDH) [171]. For addressing lactate production specifically, though many attempts have been made, these studies have been unable to significantly reduce lactate production without also affecting cell growth. A major part of this may be the use of single genes for engineering. The cell is good at maintaining homeostasis, and so a bigger disruption to gene expression may be needed to rewire glucose metabolism. To better understand this phenomenon, we previously used an optimization-based approach to identify combinations of genes that can modify glycolytic flux without interfering with certain reactions predicted to be required for growth [40].

In this study, we overexpressed genes from the malate-aspartate shuttle pathway, identified from previous optimization, into a CHO cell line producing recombinant TNFR-Fc. The engineered cell pool had similar growth to non-engineered cells but had reduced waste product generation in late stage culture. We then applied a new version of the kinetic model containing reactions for glutamine metabolism to better understand the changes made by engineering the malate aspartate shuttle.

7.2 Materials and Methods

7.2.1 Cell culture

For this study, a DXB11-based CHO cell line expressing TNFR-Fc was utilized. Cells were grown in a 37°C, 5% CO_2 incubator in 125 mL shake flasks using a chemically defined medium and were agitated on an orbital shaker at 120 rpm. Engineered cells were maintained in 6 µg/mL puromycin.

7.2.2 Vector Construction

Three gene cassette assembly of genes for metabolic engineering was performed using Golden Gate assembly as previously described [172]. Human coding sequences (CDS) for GOT1, SLC25A11, and SLC25A12 were modified to remove restriction enzyme recognition sites for *AarI*, *SapI*, and *BbsI* enzymes while maintaining the amino acid sequence. The three genes were assembled into vectors containing promoters, terminators, and position specific *BbsI* scars for multi-gene cassette assembly using *SapI*. *BbsI* was used for the final assembly into a transposon vector (Figure 7-1).

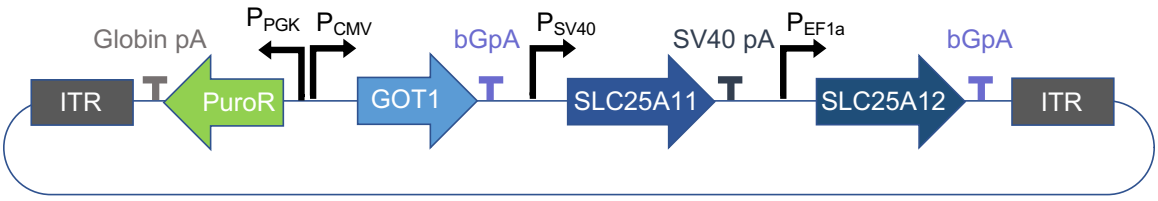

Figure 7-1. Vector for introducing metabolic gene engineering cassette into CHO cells. Vector is flanked by ITRs for the transposase and contains puromycin resistance gene for selection.

7.2.3 Cell engineering

Multi-gene cassette was integrated into CHO cell using LEAP-In Transposase (ATUM). $2x10^6$ cells were washed in PBS, re-suspended in Opti-MEM medium, and transferred to a 0.4 cm cuvette for electroporation. 4 µg of Plasmid DNA and 4 µg of transposase RNA were mixed in with the cells, followed by electroporation using a 15 msec, 300V square wave in a BioRad GenePulser Xcell. Electroporated cells were transferred to a well in a 6-well plate, and were placed in a 37°C, 5% CO_2 incubator without agitation for 24 hours. Cells were transferred to a shaker platform and further allowed to

recover. 72 hours after electroporation, the cells were seeded at 4 x 10^5 cells/mL at a 3 mL working volume in a 6-well plate with 6 µg/mL puromycin for selection. The cell pool was ready for characterization after 10 days of selection.

7.2.4 Fed-batch cultures

For fed-batch cultures, control and engineered cells (both in triplicate) were seeded at 2 x 10^5 cells/mL at a 35 mL volume in 125 mL shake flasks. Cultures were sampled daily for cell counting and measurement of key metabolites. Cell counting was performed using a Nexelcom Cellometer Auto T4. Glucose, lactate, glutamine, glutamate, and ammonia were measured on a YSI 2900D-3.

Starting on day 3, glucose, glutamine, and a concentrated feed medium were added to the cells daily. Glucose was added to increase the concentration by either 10mM (days 3-5) or 15mM (days 6-9). 1mM glutamine was added to the culture daily. Feed medium was added at 2% of culture volume on days 3 and 4, and at 3% of volume on days 5-9.

Table 7-1. qPCR primers.

Primer	Sequence
ACTB-F	AGCTGAGAGGGAAATTGTGCG
ACTB-R	GCAACGGAACCGCTCATT
GOT1-Human-F	GGATGCAGAGAAGAGAGGATTG
GOT1-Human-R	GTGGAGGACAACAATGGAGAA
GOT1-CHO-F	CAAGAACTTCGGGCTCTACAA
GOT1-CHO-R	GGGATTGGACCAAGTGATTCT
SLC25A11-Human-F	TACTGGACTCAGGCTACTTCTC
SLC25A11-Human-R	ATGTTCTGGATTCGGGTCTTG
SLC25A11-CHO-F	TGGGCTGGATGTGCTGATG
SLC25A11-CHO-R	CCAAGAAGATGAAGGTGAGGACAG
SLC25A12-Human-F	CTGTTCCACTTCCAGCAGAA
SLC25A12-Human-R	CAGCTACTTGCAGACGAATCT
SLC25A12-CHO-F	GGACTCTACCGAGGTCTGATAC
SLC25A12-CHO-R	ACAACCTCCAGCAAGGATTT

7.2.5 qRT-PCR for transcript quantification

RNA was extracted from 1 x 10^6 cells using a Zymo QuickRNA micro kit (Zymo Research). cDNA synthesis was performed using SuperScript III First-Strand cDNA

Synthesis SuperMix for qRT-PCR. qRT-PCR was done using SYBR Select Master Mix on a BioRad CFX Connect qPCR machine. Primers used for qPCR are in Table 7-1.

7.2.6 Enzyme-linked immunosorbent assay (ELISA)

Measurement of TNFR-Fc titer in cell culture supernatant was performed using ELISA. Nunc MaxiSorp ELISA plates were coated with Anti-Human IgG, Fc specific antibody (Sigma-Aldrich I3391, St. Louis, MO). Monoclonal Anti-Human IgG-Alkaline Phosphatase antibody (Sigma-Aldrich A2064, St. Louis, MO) was used as the secondary antibody, and the signal was detected using Sigmafast p-Nitrophenyl Phosphate tablets. Previously purified TNFR-Fc from these cells was used as the standard.

7.2.7 RNA-seq data

RNA from the original CHO cells was extracted using RNeasy Mini Kit (Qiagen, Valencia, CA). Strand-specific Illumina library preparation and sequencing on Illumina NovaSeq using 150bp paired end reads was performed at the University of Minnesota Genomics Center.

Data was trimmed to remove adapter sequences using Trimmomatic [55] (version 0.33). Mapping to the CriGri-PICR release of the Chinese hamster genome [56] was performed using STAR [173] with standard ENCODE options. Quantification of transcript levels was done using Cufflinks [174].

7.2.8 Kinetic model

A previously developed kinetic model of metabolism [40, 175] was modified to include reactions for glutamine and glutamate transport, as well as glutamine synthetase. Rate equations and kinetic constants for the glutamate/sodium transporter (*rgluna*), the glutamine/sodium transporter (*rglnna*), the glutamine/hydrogen transporter (*rglnh*), and glutamine synthetase (*rgs*), were obtained from a model of liver metabolism [176]. An additional simple, sodium-independent transporter for glutamate was also implemented. The rate of ammonia production was calculated from the rates of glutaminase (*gls*), glutamate dehydrogenase (*gdh*), and glutamine synthetase (*gs*):

$$r_{nh3} = r_{gls} + r_{gdh} - r_{gs}$$

After adding the new equations to the model, Vmax values for all of the enzymes were fit to experimental data using a previously described optimization scheme using a penalty value to eliminate changes that do not contribute to the objective [40]. Specific glucose consumption, lactate production, glutamine consumption, glutamate consumption, and ammonia production from fed-batch culture were set as the objective function for the optimization. The penalty value used to identify the model fit parameters was taken to be the highest value before significant deterioration in the model fit to reduce the overall degree of enzymatic change. This optimization problem was solved using the global optimization solver, BARON [177].

7.3 Results and Discussion

7.3.1 Malate-aspartate shuttle predicted to have impact on metabolism

In a previous work, mathematical optimization of a kinetic model was performed to predict modifications to enzyme expression levels that could reduce lactate production while maintaining fluxes predicted to be required for cell growth [40]. One of the key genes found in the optimization was glutamic-oxaloacetic transaminase 1 (GOT1), a member of the malate-aspartate shuttle (Figure 7-2A). The malate aspartate shuttle is the primary mechanism for shuttling NADH into the mitochondria, and can have a large impact on flux through glycolysis [178]. We further investigated the genes in this pathway by examining expression of three genes in the shuttle, GOT1 and the two transporters (SLC25A11 and SLC25A12), over the course of a fed-batch culture of a CHO cell line (Figure 7-2B). Interestingly, there is a decrease in the expression of GOT1 over time in culture, corresponding with a transition to the stationary phase of culture. Given the time dynamics evident among this gene between different growth and metabolic states, we pursued the malate aspartate genes as targets for cell engineering.

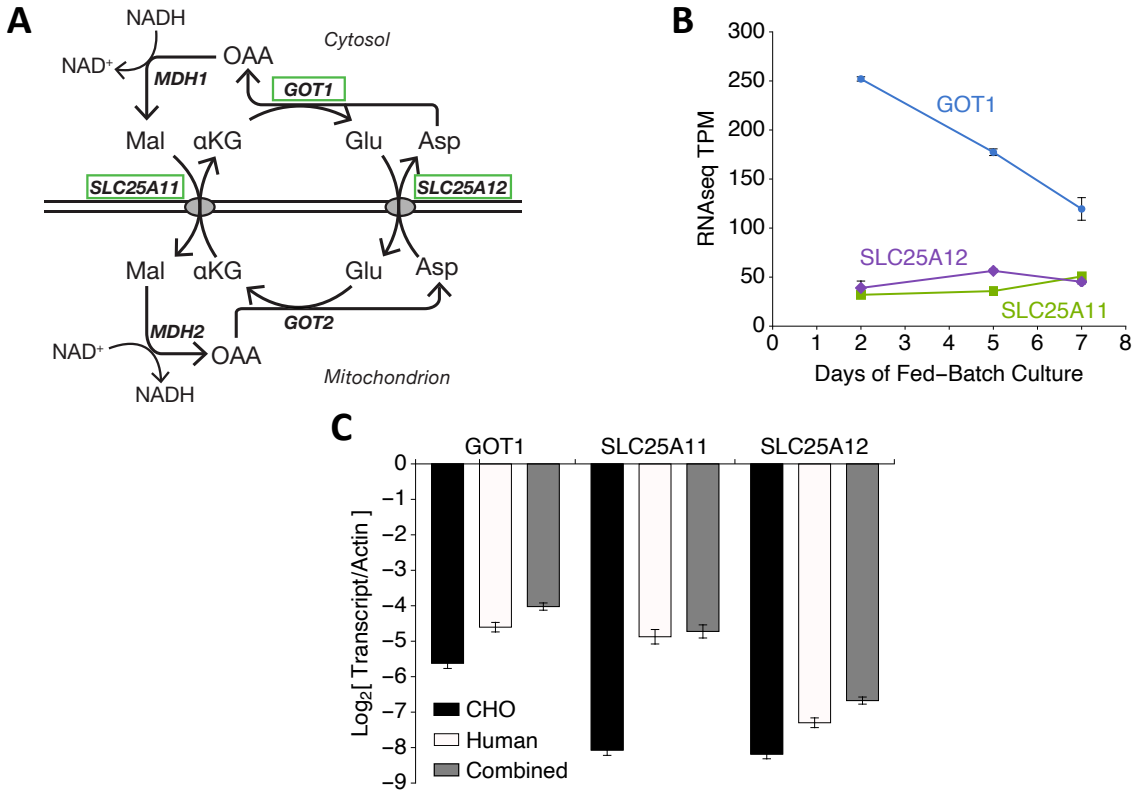

Figure 7-2. Selection of gene targets and resulting expression in engineered cells. (A) Malate-aspartate shuttle genes and pathway. Engineering targets for overexpression highlighted in green. (B) Expression of target genes during fed-batch culture in Transcripts Per Million transcripts (TPM). (C) qRT-PCR data for transcript level of target genes in engineered cell pool. Expression shown for native CHO genes, exogenously introduced human genes, and the combined expression relative to actin.

A CHO cell line producing recombinant TNFR-Fc was chosen for cell engineering. Human coding sequences for GOT1, SLC25A11, and SLC25A12 were integrated into the cells using a transposon-based system, allowing for intact integration of the multi-gene engineering cassette. The transcript level of both the native CHO and human versions of the malate aspartate genes was quantified by qRT-PCR (Figure 7-2C). GOT1 is expressed at a higher level than the transporters SLC25A11 and SLC25A12 ($2^{5.6}$-fold lower than actin as opposed to ~$2^{8.1}$ fold lower than actin). The expression of the human genes was higher in all cases than the native level in CHO. Upon calculation of the total fold increase of overall gene expression for each gene, it was determined that this system increased the overall transcript level 3-10x above the native gene level.

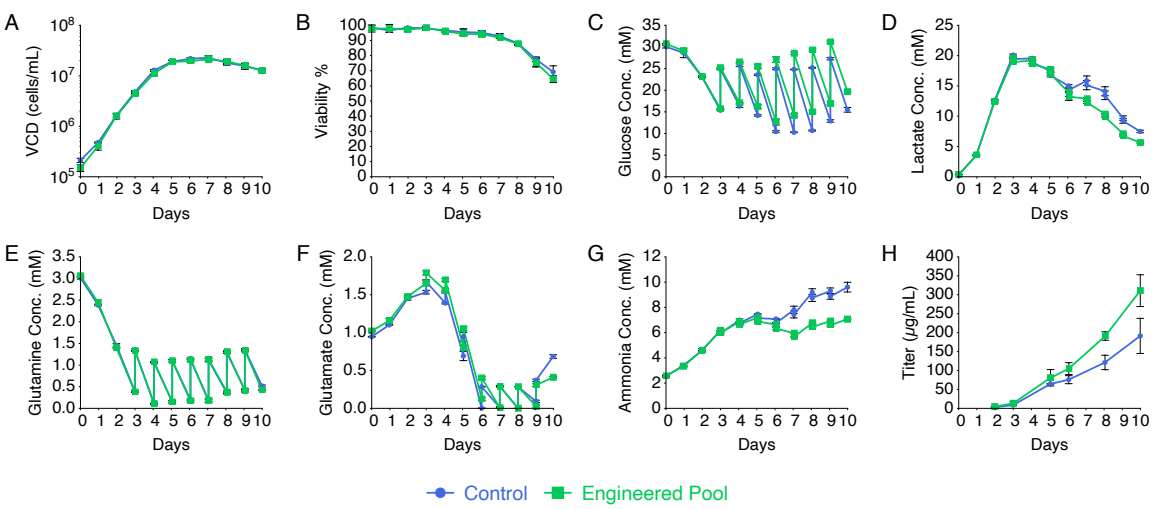

Figure 7-3. Fed-batch cultures of control cells (blue) and engineered cell pool (green). (A) Viable cell density (VCD). (B) Percent viability. (C) Glucose concentration. (D) Lactate concentration. (E) Glutamine concentration. (F) Glutamate concentration. (G) Ammonia concentration. (H) TNFR-Fc titer. N=3 replicates for each cell line/pool.

7.3.2 Engineered cell pool has reduced waste metabolite production in late stage culture

Having generated a cell pool with a reasonable level of overexpression of the target genes, fed-batch cultures were run for the control and engineered cells to characterize their behavior. During the exponential phase of the cells, their behavior is very similar. Viable cell density and viability was the same between the two cell types (Figure 7-3A, B). Under these conditions, the non-engineered cells have interesting lactate and ammonia production behavior, with a regular lactate/ammonia production state during exponential growth, a switch to consumption of both metabolites in stationary phase, and then a switch back to production for a day before the culture starts to die (Figure 7-3D, G). The engineered cells exhibit reduced lactate and ammonia production is late stage culture, by avoiding the "switch-up" to a high flux state. As for glucose, the engineered cell pool seemed to consume less glucose in the late stage of the culture, as the glucose was accumulating faster in these cells (Figure 7-3C). Glutamine and glutamate are also very similar between the two different cell lines/pool (Figure 7-3E, F). Interestingly, the titer at harvest was higher significantly higher in the engineered cell pool as opposed to the control cell line (Figure

7-3H, *t-test*, p=0.03). This could be due to the improved culture behavior from reduced waste products accumulating in the media.

On a per cell basis, overexpression of the malate aspartate rewired some specific rates in late stage culture. Specific glucose consumption is the same between the two cell types (Figure 7-4A), but specific lactate production is significantly lower in the late stage of culture in the engineered cells (Figure 7-4B). The engineered cells also have reduced specific ammonia production and increased specific glutamate consumption in late stage culture (after day 5, Figure 7-4C, D).

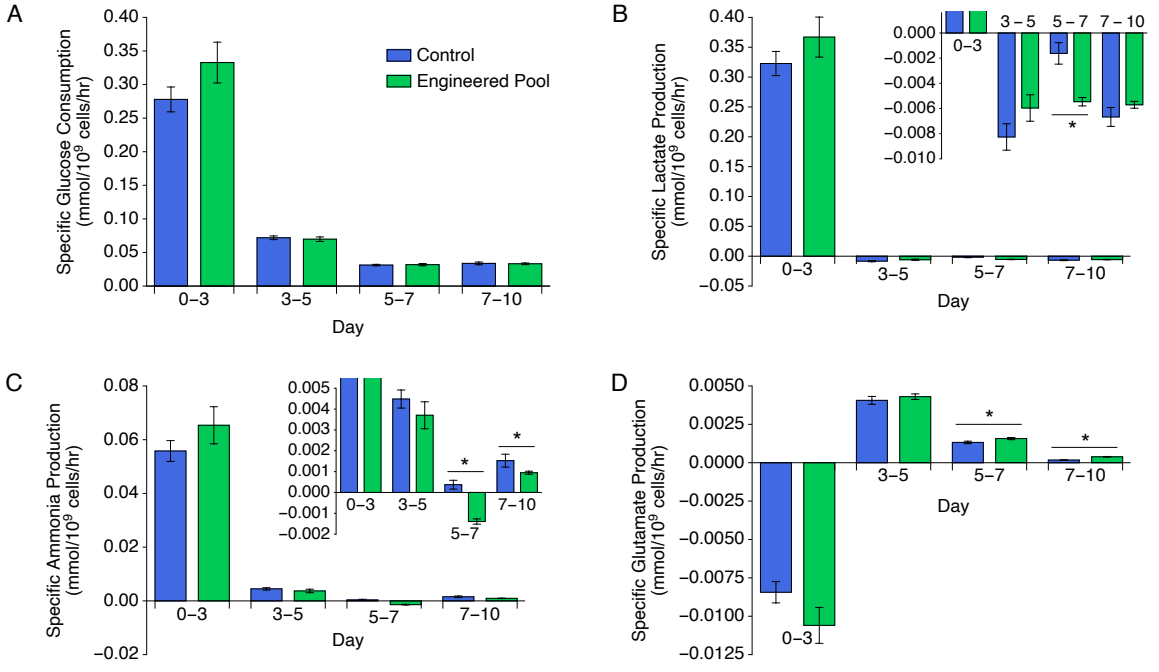

Figure 7-4. Specific rates over different phases of culture. (A) Specific glucose consumption. (B) Specific lactate production. (C) Specific ammonia production. (D) Specific glutamate consumption. (* = $p < 0.05$). Control cells are in blue, engineered cells are in green.

7.3.3 Modeling control and engineered cells

One of the major differences between the engineered and control cells was ammonia production. In order to better study this phenomenon, a previously described kinetic model of metabolism encompassing glycolysis, the pentose phosphate pathway, and the TCA cycle was modified to include reactions for glutamine metabolism. After adding the new reactions and optimizing the Vmax for each reaction to match experimental data, a

preliminary baseline model of metabolism was established (Figure 7-5). The newly optimized model has bistability in a very narrow region, and usually predicts a high flux state. Work is being done to better fit the model to experimental data and examine the different fluxes in central metabolism upon engineering the malate aspartate shuttle.

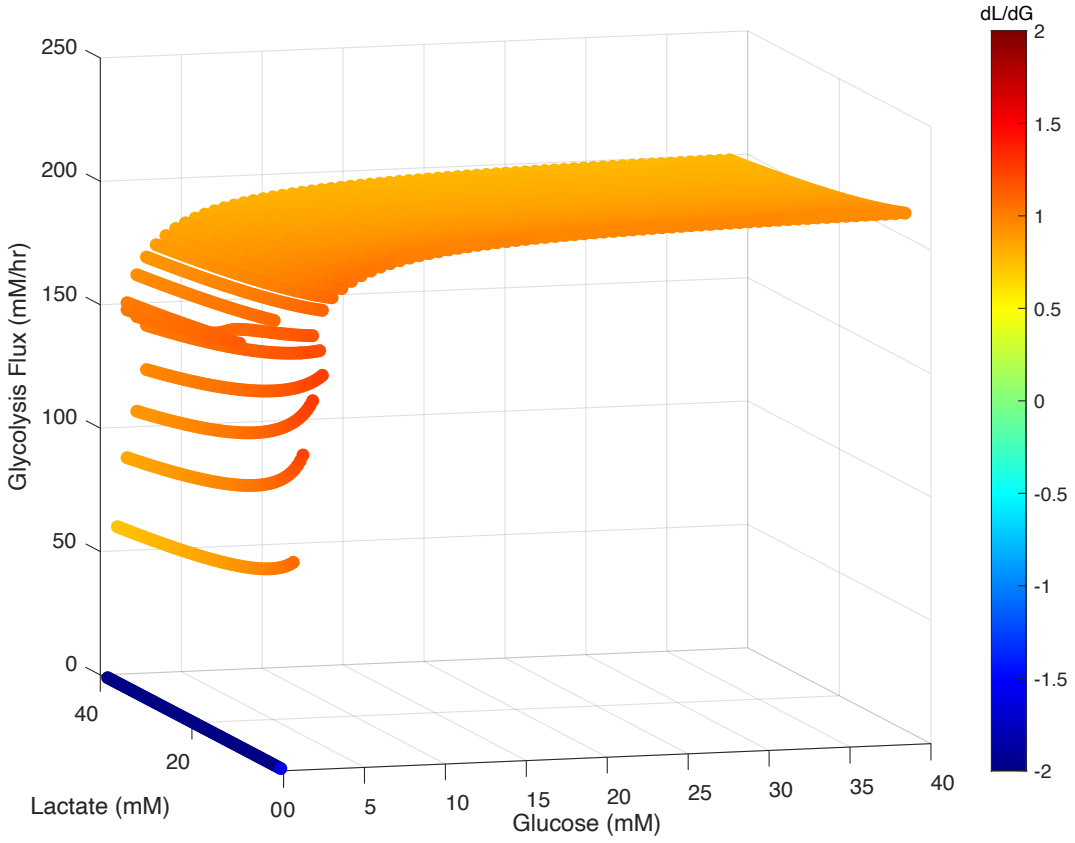

Figure 7-5. Glycolysis flux as a function of extracellular glucose and lactate concentration for new optimized model. Points are colored by dL/dG (ratio of lactate production to glucose consumption).

7.4 Conclusion

Through the use of mathematical optimization of a kinetic model, we were able to predict and implement changes to cellular metabolism that reduced waste product production. Additionally, optimization allowed us to create a cell-line specific model fit to our experimental data, which will allow us to gain greater insight into what is occurring within the cells. Use of a model guided approach will help save time and resources for engineering host cell lines in the future.

8 Summary and concluding remarks

For over 30 years, continuous improvements have been made to cell culture bioprocessing. Routine protein production is at levels that were once unimaginable. With technology constantly improving, new ways of developing cell lines are being explored. The push towards "designer cell lines" comes with the aim of shorter development times, and the ability to quickly generate ideal cell lines with minimal clone selection and stability studies. Understanding transgene integration sites and cell physiology is a large part of achieving this goal.

In this thesis, multiple aspects of the relationship between integration site and resulting cell line behavior were investigated. First, the development a rapid integration site identification method facilitates further analysis of integration sites in complex cell lines. Studying integration sites in well characterized cell lines known to have desirable qualities will aid with selecting integration sites in the future. To examine genomic instability, parental cells were compared with high and low producing subclones, leading to identification of genomic regions vulnerable to copy gain/loss. Further, a large-scale analysis across many CHO cell lines was done to look for global regions of genomic variation, independent of individual cell lines. To understand integration sites with high transcriptional potential, integration sites from high expressing cells were examined and expression was correlated to high transcriptional activity and accessibility of the integration region. All together, we can build a more complete picture of a desirable integration site, including information on stability, accessibility, and transcriptional activity.

Besides the integration site, a designer cell line would require ideal metabolic and secretory behavior, and the production of specific post-translational modifications. Metabolic behavior of CHO cells was examined through the use of a kinetic model. By engineering targets predicted by optimization of our kinetic model, waste metabolite production by the cells was reduced, showing that cellular behavior can be improved through cell line engineering. In the future, many aspects of cells can be improved via this method.

Though much of this work is targeted towards improved production of biologics, there is an opportunity to apply these types of analyses towards next generation

therapeutics, such as cell therapy. For products such as CAR-T cells, most are generated using random integration of viral vectors. By applying integration site analysis to these cells, more can be understood about the cell populations generated during development and if certain integration sites correlate to good or bad patient outcome. In the future, targeted integration will be applied to these therapies as well. Insights from this work can serve as a guide towards improved development of next generation therapies.

9 References

1. Altena, F.W. and G. Belfort, *Lateral migration of spherical particles in porous flow channels: application to membrane filtration.* Chemical Engineering Science, 1984. **39**(2): p. 343-355.
2. Walther, J., J. McLarty, and T. Johnson, *The effects of alternating tangential flow (ATF) residence time, hydrodynamic stress, and filtration flux on high-density perfusion cell culture.* Biotechnol Bioeng, 2019. **116**(2): p. 320-332.
3. Radoniqi, F., et al., *Computational fluid dynamic modeling of alternating tangential flow filtration for perfusion cell culture.* Biotechnol Bioeng, 2018. **115**(11): p. 2751-2759.
4. O'Brien, C.M., et al., *Regulation of Metabolic Homeostasis in Cell Culture Bioprocesses.* Trends in Biotechnology, 2020: p. S0167779920300354.
5. Wolf, M.K.F., et al., *Improved Performance in Mammalian Cell Perfusion Cultures by Growth Inhibition.* Biotechnol J, 2019. **14**(2): p. e1700722.
6. Yongky, A., et al., *Mechanism for multiplicity of steady states with distinct cell concentration in continuous culture of mammalian cells.* Biotechnol Bioeng, 2015. **112**(7): p. 1437-45.
7. Abu-Absi, N.R., et al., *Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe.* Biotechnol Bioeng, 2011. **108**(5): p. 1215-21.
8. Berry, B.N., et al., *Quick generation of Raman spectroscopy based in-process glucose control to influence biopharmaceutical protein product quality during mammalian cell culture.* Biotechnol Prog, 2016. **32**(1): p. 224-34.
9. O'Mara, B., et al., *Impact of depth filtration on disulfide bond reduction during downstream processing of monoclonal antibodies from CHO cell cultures.* Biotechnol Bioeng, 2019. **116**(7): p. 1669-1683.
10. Clarke, C., et al., *Transcriptomic analysis of IgG4 Fc-fusion protein degradation in a panel of clonally-derived CHO cell lines using RNASeq.* Biotechnol Bioeng, 2019. **116**(6): p. 1556-1562.
11. Wang, Q., et al., *Characterization of intact glycopeptides reveals the impact of culture media on site-specific glycosylation of EPO-Fc fusion protein generated by CHO-GS cells.* Biotechnol Bioeng, 2019. **116**(9): p. 2303-2315.
12. Le, T., et al., *An integrated platform for mucin-type O-glycosylation network generation and visualization.* Biotechnol Bioeng, 2019. **116**(6): p. 1341-1354.
13. Lavoie, R.A., et al., *Targeted capture of Chinese hamster ovary host cell proteins: Peptide ligand binding by proteomic analysis.* Biotechnol Bioeng, 2020. **117**(2): p. 438-452.

14. Chiu, J., et al., *Knockout of a difficult-to-remove CHO host cell protein, lipoprotein lipase, for improved polysorbate stability in monoclonal antibody formulations.* Biotechnol Bioeng, 2017. **114**(5): p. 1006-1015.

15. Sakuma, C., et al., *Novel endogenous simian retroviral integrations in Vero cells: implications for quality control of a human vaccine cell substrate.* Sci Rep, 2018. **8**(1): p. 644.

16. Balasubramanian, S., et al., *Generation of High Expressing Chinese Hamster Ovary Cell Pools Using the Leap-In Transposon System.* Biotechnol J, 2018. **13**(10): p. e1700748.

17. Balasubramanian, S., et al., *Comparison of three transposons for the generation of highly productive recombinant CHO cell pools and cell lines.* Biotechnol Bioeng, 2016. **113**(6): p. 1234-43.

18. O'Brien, S.A., et al., *Single Copy Transgene Integration in a Transcriptionally Active Site for Recombinant Protein Synthesis.* Biotechnol J, 2018. **13**(10): p. e1800226.

19. Zhang, L., et al., *Recombinase-mediated cassette exchange (RMCE) for monoclonal antibody expression in the commercially relevant CHOK1SV cell line.* Biotechnol Prog, 2015. **31**(6): p. 1645-56.

20. Inniss, M.C., et al., *A novel Bxb1 integrase RMCE system for high fidelity site-specific integration of mAb expression cassette in CHO Cells.* Biotechnol Bioeng, 2017. **114**(8): p. 1837-1846.

21. Carver, J., et al., *Maximizing antibody production in a targeted integration host by optimization of subunit gene dosage and position.* Biotechnol Prog, 2020: p. e2967.

22. Vcelar, S., et al., *Karyotype variation of CHO host cell lines over time in culture characterized by chromosome counting and chromosome painting.* Biotechnol Bioeng, 2018. **115**(1): p. 165-173.

23. Vcelar, S., et al., *Changes in Chromosome Counts and Patterns in CHO Cell Lines upon Generation of Recombinant Cell Lines and Subcloning.* Biotechnol J, 2018. **13**(3): p. e1700495.

24. Osada, N., et al., *The genome landscape of the african green monkey kidney-derived vero cell line.* DNA Res, 2014. **21**(6): p. 673-83.

25. Bandyopadhyay, A.A., et al., *Recurring genomic structural variation leads to clonal instability and loss of productivity.* Biotechnol Bioeng, 2019. **116**(1): p. 41-53.

26. Xiong, K., et al., *Reduced apoptosis in Chinese hamster ovary cells via optimized CRISPR interference.* Biotechnol Bioeng, 2019. **116**(7): p. 1813-1819.

27. Tang, D., et al., *Pyruvate Kinase Muscle-1 Expression Appears to Drive Lactogenic Behavior in CHO Cell Lines, Triggering Lower Viability and Productivity: A Case Study.* Biotechnol J, 2019. **14**(4): p. e1800332.

28. Mulukutla, B.C., et al., *Metabolic engineering of Chinese hamster ovary cells towards reduced biosynthesis and accumulation of novel growth inhibitors in fed-batch cultures.* Metab Eng, 2019. **54**: p. 54-68.

29. Duroy, P.O., et al., *Characterization and mutagenesis of Chinese hamster ovary cells endogenous retroviruses to inactivate viral particle release.* Biotechnol Bioeng, 2020. **117**(2): p. 466-485.

30. Dey, A.K., et al., *cGMP production and analysis of BG505 SOSIP.664, an extensively glycosylated, trimeric HIV-1 envelope glycoprotein vaccine candidate.* Biotechnol Bioeng, 2018. **115**(4): p. 885-899.

31. Kim, C.L., et al., *Improving the production of recombinant human bone morphogenetic protein-4 in Chinese hamster ovary cell cultures by inhibition of undesirable endocytosis.* Biotechnol Bioeng, 2018. **115**(10): p. 2565-2575.

32. Thoring, L., et al., *High-yield production of "difficult-to-express" proteins in a continuous exchange cell-free system based on CHO cell lysates.* Sci Rep, 2017. **7**(1): p. 11710.

33. Stolfa, G., et al., *CHO-Omics Review: The Impact of Current and Emerging Technologies on Chinese Hamster Ovary Based Bioproduction.* Biotechnol J, 2018. **13**(3): p. e1700227.

34. Kelly, P.S., et al., *Ultra-deep next generation mitochondrial genome sequencing reveals widespread heteroplasmy in Chinese hamster ovary cells.* Metab Eng, 2017. **41**: p. 11-22.

35. Berger, A., et al., *Overexpression of transcription factor Foxa1 and target genes remediates therapeutic protein production bottlenecks in Chinese hamster ovary cells.* Biotechnol Bioeng, 2020. **20**: p. 20.

36. Sumit, M., et al., *Dissecting N-Glycosylation Dynamics in Chinese Hamster Ovary Cells Fed-batch Cultures using Time Course Omics Analyses.* iScience, 2019. **12**: p. 102-120.

37. Antoniewicz, M.R., *A guide to (13)C metabolic flux analysis for the cancer biologist.* Exp Mol Med, 2018. **50**(4): p. 19.

38. Ahn, W.S., S.B. Crown, and M.R. Antoniewicz, *Evidence for transketolase-like TKTL1 flux in CHO cells based on parallel labeling experiments and (13)C-metabolic flux analysis.* Metab Eng, 2016. **37**: p. 72-78.

39. Hefzi, H., et al., *A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism.* Cell Syst, 2016. **3**(5): p. 434-443 e8.

40. O'Brien, C., et al., *Kinetic model optimization and its application to mitigating the Warburg effect through multiple enzyme alterations.* Metab Eng, 2019. **56**: p. 154-164.

41. Sewell, D.J., et al., *Enhancing the functionality of a microscale bioreactor system as an industrial process development tool for mammalian perfusion culture.* Biotechnol Bioeng, 2019. **116**(6): p. 1315-1325.

42. Sandner, V., et al., *Scale-Down Model Development in ambr systems: An Industrial Perspective.* Biotechnol J, 2019. **14**(4): p. e1700766.

43. Vormittag, P., et al., *A guide to manufacturing CAR T cell therapies.* Curr Opin Biotechnol, 2018. **53**: p. 164-181.

44. Bandyopadhyay, A.A., et al., *Advancement in bioprocess technology: parallels between microbial natural products and cell culture biologics.* J Ind Microbiol Biotechnol, 2017. **44**(4-5): p. 785-797.

45. FDA, *Points to consider in the manufacture and testing of monoclonal antibody products for human use (1997). U.S. Food and Drug Administration Center for Biologics Evaluation and Research.* J Immunother, 1997. **20**(3): p. 214-43.

46. Ochman, H., A.S. Gerber, and D.L. Hartl, *Genetic applications of an inverse polymerase chain reaction.* Genetics, 1988. **120**(3): p. 621-3.

47. Potter, C.J. and L. Luo, *Splinkerette PCR for mapping transposable elements in Drosophila.* PLoS One, 2010. **5**(4): p. e10168.

48. Schmidt, M., et al., *High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR).* Nat Methods, 2007. **4**(12): p. 1051-7.

49. de Vree, P.J., et al., *Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping.* Nat Biotechnol, 2014. **32**(10): p. 1019-25.

50. Arens, A., et al., *Bioinformatic clonality analysis of next-generation sequencing-derived viral vector integration sites.* Hum Gene Ther Methods, 2012. **23**(2): p. 111-8.

51. Giordano, F.A., et al., *High-throughput monitoring of integration site clonality in preclinical and clinical gene therapy studies.* Mol Ther Methods Clin Dev, 2015. **2**: p. 14061.

52. Aeschlimann, S.H., et al., *Enhanced CHO Clone Screening: Application of Targeted Locus Amplification and Next-Generation Sequencing Technologies for Cell Line Development.* Biotechnol J, 2019. **14**(7): p. e1800371.

53. Ho, S.C., et al., *IRES-mediated Tricistronic vectors for enhancing generation of high monoclonal antibody expressing CHO cell lines.* J Biotechnol, 2012. **157**(1): p. 130-9.

54. Yusufi, F.N.K., et al., *Mammalian Systems Biotechnology Reveals Global Cellular Adaptations in a Recombinant CHO Cell Line.* Cell Syst, 2017. **4**(5): p. 530-542 e6.

55. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-2120.

56. Rupp, O., et al., *A reference genome of the Chinese hamster based on a hybrid assembly strategy.* Biotechnology and Bioengineering, 2018. **115**(8): p. 2087-2100.

57. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.* arXiv preprint arXiv:1303.3997, 2013.

58. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

59. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Brief Bioinform, 2013. **14**(2): p. 178-92.

60. Wurm, F., *CHO Quasispecies—Implications for Manufacturing Processes.* Processes, 2013. **1**(3): p. 296-311.

61. Bandyopadhyay, A., et al., *Genomics and Systems Biotechnology in Biopharmaceutical Processing.* Chemical Engineering Progress, 2014. **110**(11): p. 45-50.

62. EMA, *ICH Harmonized Tripartite Guideline Q5D Quality of biotechnological products: derivation and characterisation of cell substrates used for production of biotechnological/biological products*. 1998.

63. Stepanenko, A., et al., *Step-wise and punctuated genome evolution drive phenotype changes of tumor cells.* Mutat Res, 2015. **771**: p. 56-69.

64. Wunsch, S., et al., *Phenotypically and karyotypically distinct Madin-Darby canine kidney cell clones respond differently to alkaline stress.* J Cell Physiol, 1995. **164**(1): p. 164-71.

65. Gaush, C.R., W.L. Hard, and T.F. Smith, *Characterization of an established line of canine kidney cells (MDCK).* Proc Soc Exp Biol Med, 1966. **122**(3): p. 931-5.

66. Rhim, J.S., et al., *Biological characteristics and viral susceptibility of an African green monkey kidney cell line (Vero).* Proc Soc Exp Biol Med, 1969. **132**(2): p. 670-8.

67. Bianchi, N.O. and J. Ayres, *Heterochromatin location on chromosomes of normal and transformed cells from African green monkey (Cercopithecus aethiops). DNA denaturation-renaturation method.* Exp Cell Res, 1971. **68**(2): p. 253-8.

68. Worton, R.G., C.C. Ho, and C. Duff, *Chromosome stability in CHO cells.* Somatic Cell Genetics, 1977. **3**(1): p. 27-45.

69. Deaven, L.L. and D.F. Petersen, *The chromosomes of CHO, an aneuploid Chinese hamster cell line: G-band, C-band, and autoradiographic analyses.* Chromosoma, 1973. **41**(2): p. 129-144.

70. Davies, J. and M. Reff, *Chromosome localization and gene-copy-number quantification of three random integrations in Chinese-hamster ovary cells and their amplified cell lines using fluorescence in situ hybridization.* Biotechnol Appl Biochem, 2001. **33**(Pt 2): p. 99-105.

71. Derouazi, M., et al., *Genetic characterization of CHO production host DG44 and derivative recombinant cell lines.* Biochemical and Biophysical Research Communications, 2006. **340**(4): p. 1069-1077.

72. Frye, C., et al., *Industry view on the relative importance of "clonality" of biopharmaceutical-producing cell lines.* Biologicals, 2016. **44**(2): p. 117-22.

73. Feichtinger, J., et al., *Comprehensive genome and epigenome characterization of CHO cells in response to evolutionary pressures and over time.* Biotechnol Bioeng, 2016. **113**(10): p. 2241-53.

74. Rouiller, Y., et al., *Reciprocal Translocation Observed in End-of-Production Cells of a Commercial CHO-Based Process.* PDA J Pharm Sci Technol, 2015. **69**(4): p. 540-52.

75. Ko, P., et al., *Probing the importance of clonality: Single cell subcloning of clonally derived CHO cell lines yields widely diverse clones differing in growth, productivity, and product quality.* Biotechnol Prog, 2017.

76. Kim, S.J., et al., *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, 1998. **58**(1): p. 73-84.

77. Barnes, L.M., C.M. Bentley, and A.J. Dickson, *Stability of protein production from recombinant mammalian cells.* Biotechnol Bioeng, 2003. **81**(6): p. 631-639.

78. Chusainow, J., et al., *A study of monoclonal antibody producing CHO cell lines: What makes a stable high producer?* Biotechnol Bioeng, 2009. **102**(4): p. 1182-1196.

79. Kim, M., et al., *A mechanistic understanding of production instability in CHO cell lines expressing recombinant monoclonal antibodies.* Biotechnol Bioeng, 2011. **108**(10): p. 2434-2446.

80. Yang, Y., et al., *DNA methylation contributes to loss in productivity of monoclonal antibody-producing CHO cell lines.* J Biotechnol, 2010. **147**(3-4): p. 180-5.

81. Osterlehner, A., S. Simmeth, and U. Gopfert, *Promoter methylation and transgene copy numbers predict unstable protein production in recombinant Chinese hamster ovary cell lines.* Biotechnol Bioeng, 2011. **108**(11): p. 2670-81.

82. Veith, N., et al., *Mechanisms underlying epigenetic and transcriptional heterogeneity in Chinese hamster ovary (CHO) cell lines.* BMC Biotechnol, 2016. **16**: p. 6.

83. Vishwanathan, N., et al., *A comparative genomic hybridization approach to study gene copy number variations among Chinese hamster cell lines.* Biotechnol Bioeng, 2017. **114**(8): p. 1903-1908.

84. Lewis, N.E., et al., *Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome.* Nat Biotechnol, 2013. **31**(8): p. 759-65.

85. Vishwanathan, N., et al., *Augmenting Chinese hamster genome assembly by identifying regions of high confidence.* Biotechnol J, 2016. **11**(9): p. 1151-7.

86. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

87. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

88. Rausch, T., et al., *DELLY: structural variant discovery by integrated paired-end and split-read analysis.* Bioinformatics, 2012. **28**(18): p. i333-i339.

89. Olshen, A.B., et al., *Circular binary segmentation for the analysis of array-based DNA copy number data.* Biostatistics, 2004. **5**(4): p. 557-72.

90. Venkatraman, E.S. and A.B. Olshen, *A faster circular binary segmentation algorithm for the analysis of array CGH data.* Bioinformatics, 2007. **23**(6): p. 657-63.

91. De Leon Gatti, M., et al., *Comparative transcriptional analysis of mouse hybridoma and recombinant Chinese hamster ovary cells undergoing butyrate treatment.* J Biosci Bioeng, 2007. **103**(1): p. 82-91.

92. Kantardjieff, A., et al., *Transcriptome and proteome analysis of Chinese hamster ovary cells under low temperature and butyrate treatment.* J Biotechnol, 2010. **145**(2): p. 143-59.

93. Yee, J.C., et al., *Genomic and proteomic exploration of CHO and hybridoma cells under sodium butyrate treatment.* Biotechnol Bioeng, 2008. **99**(5): p. 1186-204.

94. Munro, T.P., et al., *Accelerating patient access to novel biologics using stable pool-derived product for non-clinical studies and single clone-derived product for clinical studies.* Biotechnol Prog, 2017. **33**(6): p. 1476-1482.

95. Martinet, D., et al. *Karyotype of CHO DG44 cells*. in *Cell Technology for Cell Products*. 2007. Dordrecht: Springer Netherlands.

96. Cao, Y., et al., *Fluorescence in situ hybridization using bacterial artificial chromosome (BAC) clones for the analysis of chromosome rearrangement in Chinese hamster ovary cells.* Methods, 2012. **56**(3): p. 418-423.

97. Baik, J.Y. and K.H. Lee, *A framework to quantify karyotype variation associated with CHO cell line instability at a single-cell level.* Biotechnol Bioeng, 2017. **114**(5): p. 1045-1053.

98. Cao, Y., et al., *Construction of BAC-based physical map and analysis of chromosome rearrangement in Chinese hamster ovary cell lines.* Biotechnol Bioeng, 2012. **109**(6): p. 1357-67.

99. Zhou, H., et al., *Generation of stable cell lines by site-specific integration of transgenes into engineered Chinese hamster ovary strains using an FLP-FRT system.* J Biotechnol, 2010. **147**(2): p. 122-9.

100. Kim, S.J. and G.M. Lee, *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.* Biotechnol Bioeng, 1999. **64**(6): p. 741-9.

101. Jiang, Z., Y. Huang, and S.T. Sharfstein, *Regulation of recombinant monoclonal antibody production in chinese hamster ovary cells: a comparative study of gene copy number, mRNA level, and protein expression.* Biotechnol Prog, 2006. **22**(1): p. 313-8.

102. Kim, N.S., T.H. Byun, and G.M. Lee, *Key determinants in the occurrence of clonal variation in humanized antibody expression of cho cells during dihydrofolate reductase mediated gene amplification.* Biotechnol Prog, 2001. **17**(1): p. 69-75.

103. Pallavicini, M.G., et al., *Effects of methotrexate on transfected DNA stability in mammalian cells.* Mol Cell Biol, 1990. **10**(1): p. 401-4.

104. Bailey, L.A., et al., *Determination of Chinese hamster ovary cell line stability and recombinant antibody expression during long-term culture.* Biotechnol Bioeng, 2012. **109**(8): p. 2093-103.

105. Cristea, S., et al., *In vivo cleavage of transgene donors promotes nuclease-mediated targeted integration.* Biotechnol Bioeng, 2013. **110**(3): p. 871-80.

106. Lee, J.S., et al., *Accelerated homology-directed targeted integration of transgenes in Chinese hamster ovary cells via CRISPR/Cas9 and fluorescent enrichment.* Biotechnol Bioeng, 2016. **113**(11): p. 2518-23.

107. Lee, J.S., et al., *Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway.* Sci Rep, 2015. **5**: p. 8572.

108. Quigley, D.A., et al., *Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer.* Cell, 2018. **174**(3): p. 758-769 e9.

109. Xu, X., et al., *The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line.* Nat Biotechnol, 2011. **29**(8): p. 735-41.

110. Kaas, C.S., et al., *Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy.* BMC Genomics, 2015. **16**: p. 160.

111. Layer, R.M., et al., *LUMPY: a probabilistic framework for structural variant discovery.* Genome Biol, 2014. **15**(6): p. R84.

112. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

113. Chepelev, I., et al., *Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization.* Cell Res, 2012. **22**(3): p. 490-503.

114. Lagasse, H.A., et al., *Recent advances in (therapeutic protein) drug development.* F1000Research, 2017. **6**: p. 113.

115. Wurm, F., *CHO Quasispecies - Implications for Manufacturing Processes.* Processes, 2013. **1**: p. 296-311.

116. Kim, N.S., S.J. Kim, and G.M. Lee, *Clonal variability within dihydrofolate reductase-mediated gene amplified Chinese hamster ovary cells: stability in the absence of selective pressure.* Biotechnol Bioeng, 1998. **60**(6): p. 679-88.

117. Wirth, D., et al., *Road to precision: recombinase-based targeting technologies for genome engineering.* Current Opinion in Biotechnology, 2007. **18**: p. 411-419.

118. Vishwanathan, N., et al., *Transcriptome dynamics of transgene amplification in Chinese hamster ovary cells.* Biotechnol Bioeng, 2014. **111**(3): p. 518-28.

119. Le, H., et al., *Cell line development for biomanufacturing processes: recent advances and an outlook.* Biotechnol Lett, 2015. **37**(8): p. 1553-64.

120. Joung, J.K. and J.D. Sander, *TALENs: a widely applicable technology for targeted genome editing.* Nat Rev Mol Cell Biol, 2013. **14**(1): p. 49-55.

121. Sander, J.D. and J.K. Joung, *CRISPR-Cas systems for editing, regulating and targeting genomes.* Nat Biotechnol, 2014. **32**(4): p. 347-55.

122. Urnov, F.D., et al., *Genome editing with engineered zinc finger nucleases.* Nat Rev Genet, 2010. **11**(9): p. 636-46.

123. Turan, S., et al., *Recombinase-mediated cassette exchange (RMCE) - a rapidly-expanding toolbox for targeted genomic modifications.* Gene, 2013. **515**(1): p. 1-27.

124. Huang, Y., et al., *An efficient and targeted gene integration system for high-level antibody expression.* J Immunol Methods, 2007. **322**(1-2): p. 28-39.

125. Crawford, Y., et al., *Fast identification of reliable hosts for targeted cell line development from a limited-genome screening using combined phiC31 integrase and CRE-Lox technologies.* Biotechnol Prog, 2013. **29**(5): p. 1307-15.

126. Inao, T., et al., *Improved transgene integration into the Chinese hamster ovary cell genome using the Cre-loxP system.* Journal of Bioscience and Bioengineering, 2015. **120**: p. 99-106.

127. Lauth, M., *Stable and efficient cassette exchange under non-selectable conditions by combined use of two site-specific recombinases.* Nucleic Acids Research, 2002. **30**: p. e115.

128. Lauth, M., et al., *Characterization of Cre-mediated cassette exchange after plasmid microinjection in fertilized mouse oocytes.* Genesis, 2000. **27**: p. 153-158.

129. Anderson, R.P., E. Voziyanova, and Y. Voziyanov, *Flp and Cre expressed from Flp-2A-Cre and Flp-IRES-Cre transcription units mediate the highest level of dual recombinase-mediated cassette exchange.* Nucleic Acids Res, 2012. **40**(8): p. e62.

130. Koduri, R.K., J.T. Miller, and P. Thammana, *An efficient homologous recombination vector pTV(I) contains a hot spot for increased recombinant protein expression in Chinese hamster ovary cells.* Gene, 2001. **280**(1-2): p. 87-95.

131. Mielke, C., et al., *Anatomy of highly expressing chromosomal sites targeted by retroviral vectors.* Biochemistry, 1996. **35**(7): p. 2239-52.

132. Akhtar, W., et al., *Chromatin position effects assayed by thousands of reporters integrated in parallel.* Cell, 2013. **154**(4): p. 914-27.

133. Ding, S., et al., *Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice.* Cell, 2005. **122**(3): p. 473-83.

134. Hackett, C.S., A.M. Geurts, and P.B. Hackett, *Predicting preferential DNA vector insertion sites: implications for functional genomics and gene therapy.* Genome Biol, 2007. **8 Suppl 1**: p. S12.

135. Schroder, A.R., et al., *HIV-1 integration in the human genome favors active genes and local hotspots.* Cell, 2002. **110**(4): p. 521-9.

136. Wu, X. and S.M. Burgess, *Integration target site selection for retroviruses and transposable elements.* Cell Mol Life Sci, 2004. **61**(19-20): p. 2588-96.

137. Oberbek, A., et al., *Generation of stable, high-producing CHO cell lines by lentiviral vector-mediated gene transfer in serum-free suspension culture.* Biotechnol Bioeng, 2011. **108**(3): p. 600-10.

138. Matasci, M., et al., *The PiggyBac transposon enhances the frequency of CHO stable cell line generation and yields recombinant lines with superior productivity and stability.* Biotechnol Bioeng, 2011. **108**(9): p. 2141-50.

139. Li, B., M. Carey, and J.L. Workman, *The role of chromatin during transcription.* Cell, 2007. **128**(4): p. 707-19.

140. Tsompana, M. and M.J. Buck, *Chromatin accessibility: a window into the genome.* Epigenetics & Chromatin, 2014. **7**: p. 33.

141. Wallrath, L.L. and S.C.R. Elgin, *Position Effect Variegation in Drosophila Is Associated with an Altered Chromatin Structure.* Genes & Development, 1995. **9**(10): p. 1263-1277.

142. Cui, K. and K. Zhao, *Genome-Wide Approaches to Determining Nucleosome Occupancy in Metazoans Using MNase-Seq*, in *Chromatin Remodeling: Methods and Protocols*, R.H. Morse, Editor. 2012, Humana Press: Totowa, NJ. p. 413-419.

143. Song, L. and G.E. Crawford, *DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells.* Cold Spring Harb Protoc, 2010. **2010**(2): p. pdb prot5384.

144. Giresi, P.G., et al., *FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin.* Genome Res, 2007. **17**(6): p. 877-85.

145. Buenrostro, J.D., et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.* Nat Methods, 2013. **10**(12): p. 1213-8.

146. Buenrostro, J.D., et al., *ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide.* Curr Protoc Mol Biol, 2015. **109**: p. 21.29.1-9.

147. Szymczak, A.L., et al., *Correction of multi-gene deficiency in vivo using a single 'self-cleaving' 2A peptide-based retroviral vector.* Nat Biotechnol, 2004. **22**(5): p. 589-94.

148. Turan, S., et al., *Recombinase-mediated cassette exchange (RMCE): traditional concepts and current challenges.* J Mol Biol, 2011. **407**(2): p. 193-221.

149. Gibson, D.G., et al., *Enzymatic assembly of DNA molecules up to several hundred kilobases.* Nat Methods, 2009. **6**(5): p. 343-5.

150. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

151. Feng, J., et al., *Identifying ChIP-seq enrichment using MACS.* Nature Protocols, 2012. **7**(9): p. 1728-40.

152. Feller, W., *On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions.* Ann. Math. Statist., 1948. **19**(2): p. 177-189.

153. Li, X., et al., *Generation of Destabilized Green Fluorescent Protein as a Transcription Reporter.* Journal of Biological Chemistry, 1998. **273**: p. 34970-34975.

154. Cockerill, P.N., *Structure and function of active chromatin and DNase I hypersensitive sites.* FEBS J, 2011. **278**(13): p. 2182-210.

155. Shlyueva, D., G. Stampfel, and A. Stark, *Transcriptional enhancers: from properties to genome-wide predictions.* Nat Rev Genet, 2014. **15**(4): p. 272-86.

156. Whyte, W.A., et al., *Master transcription factors and mediator establish super-enhancers at key cell identity genes.* Cell, 2013. **153**(2): p. 307-19.

157. Fahrner, J.A. and S.B. Baylin, *Heterochromatin: stable and unstable invasions at home and abroad.* Genes Dev, 2003. **17**(15): p. 1805-12.

158. Ghirlando, R., et al., *Chromatin domains, insulators, and the regulation of gene expression.* Biochim Biophys Acta, 2012. **1819**(7): p. 644-51.

159. Dickson, A.J., *Importance of Genetic Environment for Recombinant Gene Expression*, in *Cell Line Development*, M. Al-Rubeai, Editor. 2009, Springer Netherlands: Dordrecht. p. 83-96.

160. Wang, X., et al., *Cre-Mediated Transgene Integration in Chinese Hamster Ovary Cells Using Minicircle DNA Vectors.* Biotechnol J, 2018: p. e1800063.

161. Mayrhofer, P., et al., *Accurate comparison of antibody expression levels by reproducible transgene targeting in engineered recombination-competent CHO cells.* Appl Microbiol Biotechnol, 2014. **98**(23): p. 9723-33.

162. Barnes, L.M., N. Moy, and A.J. Dickson, *Phenotypic variation during cloning procedures: analysis of the growth behavior of clonal cell lines.* Biotechnol Bioeng, 2006. **94**(3): p. 530-7.

163. Davies, S.L., et al., *Functional heterogeneity and heritability in CHO cell populations.* Biotechnol Bioeng, 2013. **110**(1): p. 260-74.

164. Seth, G., et al., *In pursuit of a super producer-alternative paths to high producing recombinant mammalian cells.* Curr Opin Biotechnol, 2007. **18**(6): p. 557-64.

165. Kallehauge, T.B., et al., *Ribosome profiling-guided depletion of an mRNA increases cell growth rate and protein secretion.* Scientific Reports, 2017. **7**: p. 40388.

166. Le, H., et al., *Multivariate analysis of cell culture bioprocess data--lactate consumption as process indicator.* J Biotechnol, 2012. **162**(2-3): p. 210-23.

167. Templeton, N., et al., *The impact of anti-apoptotic gene Bcl-2 expression on CHO central metabolism.* Metab Eng, 2014. **25**: p. 92-102.

168. Wilkens, C.A. and Z.P. Gerdtzen, *Comparative metabolic analysis of CHO cell clones obtained through cell engineering, for IgG productivity, growth and cell longevity.* PLoS One, 2015. **10**(3): p. e0119053.

169. Yip, S.S., et al., *Complete knockout of the lactate dehydrogenase A gene is lethal in pyruvate dehydrogenase kinase 1, 2, 3 down-regulated CHO cells.* Mol Biotechnol, 2014. **56**(9): p. 833-8.

170. Zhou, M., et al., *Decreasing lactate level and increasing antibody production in Chinese Hamster Ovary cells (CHO) by reducing the expression of lactate dehydrogenase and pyruvate dehydrogenase kinases.* J Biotechnol, 2011. **153**(1-2): p. 27-34.

171. Jeon, M.K., D.Y. Yu, and G.M. Lee, *Combinatorial engineering of ldh-a and bcl-2 for reducing lactate production and improving cell growth in dihydrofolate reductase-deficient Chinese hamster ovary cells.* Appl Microbiol Biotechnol, 2011. **92**(4): p. 779-90.

172. Stach, C.S., et al., *Model-Driven Engineering of N-Linked Glycosylation in Chinese Hamster Ovary Cells.* ACS Synth Biol, 2019. **8**(11): p. 2524-2535.

173. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

174. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.

175. Mulukutla, B.C., et al., *Multiplicity of steady states in glycolysis and shift of metabolic state in cultured mammalian cells.* PLoS One, 2015. **10**(3): p. e0121561.

176. Berndt, N., et al., *HEPATOKIN1 is a biochemistry-based model of liver metabolism for applications in medicine and pharmacology.* Nat Commun, 2018. **9**(1): p. 2386.

177. Tawarmalani, M. and N.V. Sahinidis, *A polyhedral branch-and-cut approach to global optimization.* Mathematical Programming, 2005. **103**(2): p. 225-249.

178. O'Brien, C.M., et al., *Regulation of Metabolic Homeostasis in Cell Culture Bioprocesses.* Trends in Biotechnology, 2020.