CHINESE HAMSTER OVARY CELL-SPECIFIC BIOPHARMACEUTICAL GLYCOFORM PREDICTIONS THROUGH DISCRETIZED REACTION NETWORK MODELING

by

Benjamin G. Kremkow

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemical Engineering

Fall 2016

© 2016 Benjamin G. Kremkow
All Rights Reserved

ProQuest Number: 10246137

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

ProQuest 10246137

Published by ProQuest LLC (2017). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

CHINESE HAMSTER OVARY CELL-SPECIFIC BIOPHARMACEUTICAL GLYCOFORM PREDICTIONS THROUGH DISCRETIZED REACTION NETWORK MODELING

by

Benjamin G. Kremkow

Approved: _____
Abraham M. Lenhoff, Ph.D.
Chair of the Department of Chemical and Biomolecular Engineering

Approved: _____
Babatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____
Ann L. Ardis, Ph.D.
Senior Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Kelvin H. Lee, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Babatunde A. Ogunnaike, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Maciek R. Antoniewicz, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Anne S. Robinson, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

I would like to recognize the people who supported me throughout my PhD program. I thank Dr. Kelvin Lee for his support and guidance of both my research project and myself as a researcher. I am blessed to have been advised by a leading researcher within the Chemical Engineering field who taught me how to think critically and also valued personal skill development and scientific outreach. I thank my committee members including Dr. Babatunde Ogunnaike, Dr. Maciek Antoniewicz, and Dr. Anne Robinson for their support of my dissertation research through modeling and experimental insights. I have developed as a person as well as a research scientist under your collective teachings and support. I thank both past and present Lee group members, in particular Stephanie Hammond, Diane Wuest, Kristin Valente, Amalie Levy, Jennifer Mantle, Lie Min, Jongyoun Baik, Melissa St. Amand, Devesh Radhakrishnan, Xiolin Zhang, Maddy MacDonald, and John Ruano-Salguero, for their paper and poster reviews, sharing of technique protocols, and technical insights and conversations. I thank my UD classmates, in particular Kathy Whitaker, Alan Fast, Bryan Yonemoto, Devesh Radhakrishnan, and Jill Emerson for their encouragement, assistance with technical classes, and friendship. I also thank Leila Choe for her assistance with experimental techniques, organization with all documents and orders, and coordination and support with the mass spectrometry experiments. I thank the DBI personnel who helped in the success of my research including Dr. Yu-Sung Wu and the Lenhoff research group for their assistance with chromatography column development; Shawn Polson, Karol Miaskiewicz, and Eric Garrison for their

assistance with and support of CHOgenome.org; and Erin Bernberg and the DBI Sequencing Center for their assistance with sequencing experiments. As part of my overall PhD program, I interacted with many people in Colburn Lab and DBI such as Kathie Young, Megan Argoe, Katie Lakofsky, Allie Sethman, Catherine Stoner, and the Colburn and DBI staff, and I thank all of them for their support, help, and collaboration. I have enjoyed working in the Lee Lab at DBI in part due to the supportive and focused environment. Finally, I thank my caring family and friends for their faith and support through this journey striving to achieve one of my dreams. First, I thank my wife, Jenna Kremkow, for moving to the East Coast with me to pursue our educational goals, all while growing our relationship and having a blast exploring the East Coast's sights, sounds, and most of all, food. I thank my parents, Jim and Sue Kremkow, as you have not wavered in your support of me through the struggles and successes during this journey. I thank my brothers, Andrew and Nick Kremkow, as you both are inspirations to me to continue to work hard, have a little fun along the way, and have also been a source of support, laughter, and love when I most needed it. I thank my grandma, Joan Kremkow, as you have been truly supportive of my goals through the late-night phone calls, care packages, sports updates while I was in class and studying with no ability to watch, and endless ideas of meals for one. I thank the rest of my extended family and the Davis family for their confidence and support in all of my endeavors. Friends who have specifically supported me during my PhD program include new Delaware friends such as Lyndsey Fisher, Mary Watermann, Sam Ferrara, musicians within the First State Symphonic Band, and Renee Henry as well as friends from Michigan including Jeannie Klavanian, Heather Smith, Nicole VanderZouwen, Adam Loyson, and the other

Spartan ChemE's who also successfully obtained a PhD at universities across the country. I am grateful to everyone for their assistance, generosity, and support during my successful journey in obtaining my PhD.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Chinese hamster ovary (CHO) cells produce more biopharmaceuticals than any other cell line due to the CHO cells' many advantageous characteristics, including their ability to glycosylate biopharmaceuticals with human-compatible glycans. The biopharmaceutical glycoform affects the product efficacy, half-life, and immunogenicity; therefore, biopharmaceuticals must have a consistent glycoform to ensure therapeutic efficacy and patient safety. One challenge associated with ensuring consistent glycosylation is that the CHO-specific glycosylation reaction network is a non-template driven cellular process with many variables, making predictive glycoform modeling difficult. This research addresses this challenge through the design of a computational glycosylation tool using a novel modeling technique that predicts biopharmaceutical glycoforms and thereby generates experimentally-relevant CHO cell line-specific information to improve biopharmaceutical manufacturing.

Achieving this goal requires a fundamental understanding of CHO cellular biology, processes, and reaction networks. The recently sequenced and annotated CHO genome facilitates a detailed, mechanistic understanding of CHO cell-specific biology. Recent and emerging genome sequencing technologies were characterized, the differences between the technologies were highlighted, and the reported CHO biopharmaceutical applications of these sequencing technologies were examined with

a focus on the sequencing and annotation of the CHO-K1 cell and Chinese hamster (CH) genomes. We improved CHOgenome.org, the centralized CHO community's public database repository through the addition of the CHO and CH genomes to the website databases and through the creation of additional genomic and proteomic bioinformatics tools. The reported CHO research community's use of these bioinformatics tools was also described.

Controllability of the biopharmaceutical product quality is essential for an approved biotherapeutic and predicting the results of cell-engineered modifications on the glycoform could aid future product quality control methods. This work details the use of a novel Discretized Reaction Network Modeling using Fuzzy Parameters (DReaM-zyP) modeling technique that was then used to create Glyco-Mapper, an innovative systems biology glycosylation prediction tool. The Glyco-Mapper input variables consist of glycosylation gene parameters and the media's nutrient composition, enabling Glyco-Mapper to replicate cell line-specific reference glycoforms and predict the glycoform changes resulting from various cell engineering modifications. The modifications Glyco-Mapper has successfully predicted include the altered expression of glycosylation, nucleotide sugar transport, and metabolism genes, as well as modified nutrient feeding strategies. Glyco-Mapper's ability to replicate cell line-specific reference glycoforms and accurately predict the reference-specific engineered glycoforms provides a streamlined tool to design cell lines to control specific product quality attributes.

In this work, CHO-produced biopharmaceutical glycoforms from literature were used to validate the Glyco-Mapper's predictive glycoform output. Glyco-Mapper predicted the engineered glycoforms with an accuracy, sensitivity, specificity, and predictive accuracy of the glycans changing experimental measurements as a result of the engineering strategy of 96%, 85%, 97%, and 85%, respectively. A non-mAb model biopharmaceutical reference glycoform was replicated using Glyco-Mapper, a novel gene knockdown (*GnT-II*) was predicted, and the predicted glycoform was experimentally confirmed with an accuracy and specificity of 95% and 98%, respectively. Additional glycoform predictions are presented and continued investigation of these predictions is recommended to further validate and improve the Glyco-Mapper tool. Glyco-Mapper is a novel CHO-specific glycosylation tool that predicts biopharmaceutically-relevant glycoforms and generates CHO cell line-specific information that can be used to improve biopharmaceutical manufacturing through enhanced product quality control.

Chapter 1

INTRODUCTION

1.1 Project Background and Motivation

1.1.1 Glycoprofile Impact on Biopharmaceutical Manufacturing

A biopharmaceutical is a therapeutic macromolecule or cell, most commonly a protein produced in a genetically modified host organism, used to treat diseases, including oncological and immunological disorders (Walsh 2010). Biopharmaceuticals on the market in 2015 numbered more than 130 and had global sales of $154 billion (La Merie 2016), an increase of $29 billion from 2012 (Kremkow and Lee 2013). Roughly 7,000 biopharmaceuticals were in the 2015 global biopharmaceutical development pipeline (PhRMA 2015) and while many of these biopharmaceuticals will fail to be approved or even tested in clinical trials, a few will meet the rigorous requirements necessary to treat human diseases. For example, only seven of the 103 biopharmaceuticals to treat melanoma that have entered the pipeline have been approved since 1998 (PhRMA 2015). Approved biopharmaceuticals must be manufactured on a large scale and as they are orders of magnitude larger and more complex than small molecule pharmaceuticals, cellular machinery is required for their production instead of chemical synthesis (Walsh 2007).

Various host organisms can be genetically modified to recombinantly express a biopharmaceutical. Mammalian cells were used to produce the majority of

biopharmaceuticals in 2015, 65% of all biotherapeutics encompassing nearly 75% of the global sales (La Merie 2016). One specific type of mammalian cell, Chinese hamster ovary (CHO) cells, are the primary biopharmaceutical expression system accounting for roughly 75% of the mammalian-produced biopharmaceuticals and global sales (La Merie 2016). The prominence of CHO cells is due to their ability to sustain high cell viability, resist viral infection, sustain high cellular biopharmaceutical productivity, and perform appropriate post-translational modifications (Jayapal et al. 2007, Griffin et al. 2007, Xu et al. 2011). Regarding the post-translational modifications, CHO cells achieve a glycosylation profile most similar to humans. One reason for the high degree of similarity is the annotated CHO genome contains homologs for all but three of the hundreds of glycosylation genes within the human genome (Xu et al. 2011).

Glycosylation is the attachment of a carbohydrate to another organic molecule's functional group, often an amino acid residue within a protein sequence. Glycosylation is divided into the five following broad classifications: N-glycosylation, O-glycosylation, glycosphingolipids, glycosaminoglycans, and glycosylphosphatidylinositol-anchored proteins (Brooks et al. 2002). N-glycosylation and O-glycosylation are the classifications commonly associated with CHO-produced biopharmaceuticals (Walsh and Jefferis 2006). N-glycosylation requires an exposed consensus amino acid sequence unlike O-glycosylation, employs experimental methodologies that are currently greater in variety and application than O-glycosylation, and is the only glycosylation classification affiliated with monoclonal antibodies (mAbs), a class of CHO-produced biotherapeutics accounting for more than half of the CHO-produced biopharmaceutical sales (La Merie 2016). For these

reasons, N-glycosylation will be the classification of reference for the remainder of this thesis. Biologically, the N-glycosylation reaction network begins with a thirteen nucleotide carbohydrate that is attached to the protein at a specific three amino acid sequence in the endoplasmic reticulum (ER). As the protein moves from the ER through the Golgi, glycosidase enzymes trim terminal nucleotides and glycosyltransferase enzymes attach additional nucleotide sugars. Once the protein leaves the Golgi, the carbohydrate altering process is generally complete. Understanding this reaction network is necessary as a biopharmaceutical's glycans affect the patient's interaction with the biopharmaceutical.

A direct link has been identified between many glycan characteristics (nucleotide composition, linkages, and carbohydrate structure) and the biopharmaceutical product quality, including the half-life, immunogenicity, and efficacy, making biopharmaceutical production glycosylation controllability critical. The biopharmaceutical half-life is decreased when a galactose (Gal) nucleotide is terminal because the liver recognizes and removes proteins containing glycans with exposed Gal nucleotides (Ashwell et al. 1974). The immunogenicity increases when the carbohydrate contains foreign (non-human) linkages or nucleotides, including the Gal-α-Gal linkage coded for by the *Ggta* gene within the CHO genome (Bosques et al. 2010). Efficacy is directly correlated with fucosylation, the presence or absence of the fucose nucleotide, for mAbs with an antibody-dependent cellular cytotoxicity (ADCC) functionality (Shinkawa et al. 2003). The effects of a biopharmaceutical's product quality, including those influenced by the glycan characteristics, upon a patient are important to understand and control.

Controllability of the glycosylation process is difficult because glycosylation is a non-template driven process, resulting in a distribution of glycans attached to the protein, not one glycan. This glycan distribution is referred to as a glycoform or glycoprofile and is one biopharmaceutical component regulated by the United States Food and Drug Administration (US DHHS FDA 2015). The biopharmaceutical glycoform composition must remain constant to ensure consistent product quality for patients. Controllability of the glycosylation process is improving through the public availability of the sequenced CHO and Chinese hamster (CH) genomes, targeted gene-editing techniques (i.e. CRISPR, TALEN), and improved glycosylation models and mass spectra analysis. However, controllability is still severely limited because the non-template driven glycosylation process is dependent upon many variables, including enzyme expression, activity, and localization as well as substrate availability. Variables have been individually altered to determine their effect (Chen and Harcum 2006; Hossler et al. 2014; Nyberg et al. 1999; Kochanowski et al. 2008; Yang et al. 2015), but an understanding of the true interconnectivity of all the variables is incomplete, making controllability difficult without further study.

1.1.2 Glycomics

Glycomics describes the large-scale study of a cell's or tissue's entire carbohydrate composition, including post-translational modification glycan carbohydrates (Brooks et al. 2002). In addition to the glycosylation process being non-template driven, glycans are often branched molecules and rarely linear. The resulting variety of complex glycan structures currently prevents one single method from comprehensively measuring the glycan characteristics. Rather, glycomics

encompasses a large set of methods defined by the various glycan characteristic they are best suited to measure, including glycoprotein confirmation and glycan identification, quantification, and characterization (Brooks et al. 2002).

Glycoprotein confirmation, verifying a protein is glycosylated, is commonly achieved with the use of sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and a glycoprotein stain (Zhou et al. 2014). Upon glycan confirmation, glycan identification is often pursued and two categories of glycan identification methods exist. One category identifies a single glycan; whereas, the other category identifies multiple glycans within the protein's glycoform. Single glycan identification often involves lectins, which are proteins or glycoproteins that bind with high specificity and selectivity to glycan carbohydrates (Lis and Sharon 1998). The high specificity of lectins enables the measurement or purification of one monosaccharide or complex oligosaccharide, depending upon the lectin in use (Tateno et al. 2007). Identification of the glycans attached to a glycoprotein identified in an SDS-PAGE gel is commonly achieved by western blotting, using lectin or antibody probing. However, additional lectins or other techniques are required to obtain information regarding the remainder of the glycoform (Tateno et al. 2007).

The identification of multiple glycans within a glycoform can occur via many methods, depending upon the level of detailed glycan data to be obtained (Sandra et al. 2014), but all methods require the initial chemical or enzymatic glycan release from the protein (O'Neill 1996). One common identification method is oligosaccharide separation and mapping by liquid chromatography, where glycan mixtures are separated by flow through high-resolution columns of various designs (Melmer et al. 2011). Electrophoresis methods are also used to separate and map glycoforms, through

either fluorophore-assisted carbohydrate electrophoresis (FACE) (Vermassen et al. 2014) or capillary electrophoresis (CE) (Mechref et al. 2005). Oligosaccharide molecular weight and charge can be measured by mass spectrometry with great accuracy, enabling the assignment of the glycan composition (Alley et al. 2013). Further information regarding the monosaccharide linkages can be determined by using gas chromatography (Zanetta et al. 2004), fast atom bombardment (Dell et al. 1991), or electrospray ionization (Prater et al. 2009), each coupled with mass spectrometry. The structure of a single, purified glycan can be determined by nuclear magnetic resonance (NMR) through measurement of the distortion of the magnetic field by the glycan (Malhotra et al. 1995). Analyzing a complete glycoform is challenging and labor-intensive, but technological advancements enable additional and more accurate biopharmaceutical glycoform data to be obtained. The supplemental data, with interpretation and analysis, will likely increase process understanding and identify or explain experimental trends. Tools, models, and software are being continuously researched, published, and updated to assist researchers with these tasks.

Within current biomanufacturing processes, glycan identification and quantification is required. Glycoform measurements within the biopharmaceutical industry are most commonly accomplished with mass spectrometry or chromatography, but each has disadvantages that should be understood (Brooks et al. 2002). Both techniques are often run "off-line," as it is difficult to generate comprehensive glycoform data in real-time. Mass spectrometry commonly requires a larger sample than most chromatography methods and only the peaks matching the glycan database defined mass/charge ratios will be identified as glycans. While chromatography requires less material, the data are not as sensitive or numerically

intensive as mass spectrometry data (Brooks et al. 2002). Measuring the glycoform composition using these glycomic methods ensures consistent biopharmaceutical glycoprofiles, yet predicting and controlling the glycoform might be easier with glycosylation enzyme activity and concentration measurements. However, the enzyme measurement assays are more difficult to perform than the glycosylation assays and the results are still obtained off-line, preventing real-time analysis. The use of glycomics methods to advance biomanufacturing process analysis is ongoing and with improved techniques, tools, software, and experimental data quality, glycosylation controllability will continue to improve.

1.1.3 Glycosylation Modeling

Computational systems biology modeling of glycosylation is required to achieve a holistic understanding of trends within experimental glycoform data. The glycosylation enzyme activities, nucleotide sugar availability, nucleotide sugar metabolism, central carbon metabolism, transport enzymes, physical cellular conditions, media feeding strategies, media supplements, enzyme localization, cellular growth rate, and cellular processing time of the biopharmaceutical each affects the final glycoform (Hossler et al. 2009, Mariño et al. 2010). Determining the influence of each of these biological factors on the glycosylation reaction network is currently too difficult to be fully captured in one comprehensive model. Due to the complex nature of glycosylation, multiple modeling approaches have been employed to better understand the glycosylation process.

The glycosylation enzymes and their respective kinetic activity levels are critical components of the glycosylation reaction network that directly affect the

biopharmaceutical glycoform. The enzyme kinetic activity levels are used in kinetic modeling to enable a model to produce quantitative glycoforms. Since 1997, multiple kinetic models have been published and continually improved upon. The recent models incorporate 19 enzymes, account for more than 10,000 glycan species, and calculate glycosylation pathway fluxes and enzyme activities (Umaña and Bailey 1997, Krambeck and Betenbaugh 2005, Krambeck et al. 2009, Liu and Neelamegham 2014). These models use differential equations and operate under the assumption that the metabolism, nucleotide sugar, and physical cellular conditions do not significantly affect the glycosylation process. One characteristic disadvantage of these kinetic models is the input requirement of an enzymatic activity level measurement that is experimentally difficult to obtain because the enzyme activity levels are likely not consistent between cell lines (Lewis et al. 2013) nor across biopharmaceutical production runs (Wong et al. 2010).

Other published glycosylation models do not solely rely on the glycosylation enzyme activity levels. One model uses Markov chains (Spahn et al. 2016) to mathematically calculate relevant matrix parameters to accurately model various glycoform distributions, not requiring user-provided kinetic information. Another publication using controllability analysis (St. Amand et al. 2014) identifies parameters affecting the specific glycoform composition in a defined pattern, specifically examining individual metal ion concentrations, media components, and cellular metabolite concentrations. Both methods enable glycosylation modeling using relevant mathematical parameters in lieu of solely experimental enzymatic activity levels. While CHO metabolic modeling is not yet linked to glycosylation modeling, experimental studies investigating correlations between glycosylation and nutrient

feeding strategies have been published (Wong et al. 2010; Maszczak-Seneczko et al. 2010). The incorporation of parameters affecting the glycosylation process in addition to enzymatic activity levels enables a more accurate representation of the glycosylation reaction network as it relates to biopharmaceutical manufacturing.

For accurate systems biology modeling, well-defined system-specific information is required. Experimental data providing this biological information is often collected using omics methods, including proteomics, transcriptomics, genomics, metabolomics, and glycomics. An annotated species genome facilitates targeted experimental data analysis and as the annotated CHO and CH genomes were released in 2011 (Xu et al. 2011) and 2013 (Lewis et al. 2013) respectively, CHO-specific gene, transcript, and protein sequences are publicly available. These annotations enable CHO-specific experimental data analysis and can provide models with numerous CHO data sets. The availability of specialized genomic and experimental data enables the creation of CHO-specific models to further advance the modeling and understanding of biopharmaceutical glycosylation.

1.2 Project Goals

The purpose of this work is to develop and validate a CHO-specific glycosylation model that predicts biopharmaceutically-relevant glycoforms and; thereby, generates CHO cell line-specific experiment design information to improve biopharmaceutical manufacturing. This goal was attained through two specific objectives:

(1) *Development of a CHO-specific glycoform prediction tool*: The CHO and CH genome annotations were systematically explored for all glycosylation-related

genes including glycosyltransferase, glycosidase, central carbon metabolism, nucleotide sugar synthesis, and nucleotide sugar transporter genes. The glycoform prediction tool, Glyco-Mapper, was created in Microsoft Excel version 2010 using the novel modeling technique Discretized Reaction Network Modeling using Fuzzy Parameters (DReaM-zyP). Glyco-Mapper incorporates multiple glycosylation parameters, including each glycosylation, metabolism, and transporter enzyme activity and the media components in a computationally simple platform to enable cell line-specific CHO glycoform predictions.

(2) *Validation of CHO-specific glycoform predictions and predictions of novel biopharmaceutically-relevant glycoform-engineering strategies*: Published CHO-specific glycoform-engineered literature was used to validate the Glyco-Mapper predictions across multiple engineering strategies. Published glycosylation, metabolic, and transporter gene overexpression and knock-out data were accurately modeled. Novel cell-engineered and medically-relevant non-mAb model biopharmaceutical glycoforms were predicted using Glyco-Mapper and one glycoform prediction was experimentally confirmed using an optimized experimental workflow.

1.3 Scope of Work

Chapter 2 identifies sequencing technologies of use to the CHO cell-based biomanufacturing community; examines recent applications of these technology platforms towards genomics, transcriptomics, proteomics, metabolomics, and multi-omics studies; and explains how next-generation sequencing methods greatly enhance omics data analysis and experiment design (Kremkow and Lee 2013). CHOgenome.org is introduced as a central CHO resource; the updated genome

annotations are described in the context of their availability on the website; the improved CHOgenome.org gene search, BLAST, and genome viewer features are described; and a few impacts CHOgenome.org has had upon the international CHO community are detailed (Kremkow et al. 2015) in Chapter 3. In Chapter 4, the glycoform prediction tool Glyco-Mapper containing more than 150 CHO-specific glycosylation-relevant genes is introduced, four literature-based glycoform-engineering predictions highlighting each engineering strategy are examined, and a novel Glyco-Mapper glycoform prediction is experimentally confirmed (Kremkow and Lee Submitted). Opportunities for additional Glyco-Mapper feature development are described and the Glyco-Mapper application of novel glycoform predictions that will further aid biopharmaceutical development are identified in Chapter 5.

In Appendix A, Sanger, second, and third generation sequencing technologies are explained and the characteristics of each are compared to identify sequencing characteristic trends regarding animal cell sequencing (Kremkow and Lee 2015). The development of the DReaM-zyP technique is outlined and the application of the DReaM-zyP technique to create Glyco-Mapper is further explained in Appendix B. Appendix C demonstrates additional literature-based Glyco-Mapper predictions, including glycoforms containing the less common GalNAc and Gal-α-Gal glycans.

REFERENCES

Allay Jr. WR, Mann BF, Novotny MV. (2013) High-sensitivity analytical approaches for the structural characterization of glycoproteins. Chem Rev. 113: 2668-2732.

Ashwell G, Morell AG. (1974) Role of surface carbohydrates in hepatic recognition and transport of circulating glycoproteins. Adv Enzymol Relat Areas Mol Biol. 41:99-128.

Bosques CJ, Collins BE, Meador JW, Sarvaiya H, Murphy JL, DelloRusso G, Bulik DA, Hsu IH, Washburn N, Sipsey SF, Myette JR, Raman R, Shriver Z, Sasisekharan R, Venkataraman G. (2010) Chinese hamster ovary cells can produce galactose-α-1,3-galactose antigens on proteins. Nat Biotechnol. 28:1153-1156.

Brooks SA, Dwek MV, Schumacher U. (2002) Functional and molecular glycobiology. BIOS Scientific Publishers Limited, Oxford, UK.

Chen P, Harcum SW. (2006) Effects of elevated ammonium on glycosylation gene expression in CHO cells. Metab Eng. 8:123-132.

Dell A, Morris HR, Greer F, Redfern JM, Rogers ME, Weisshaar G, Hiyama J, Renwick AGC. (1991) Fast-atom-bombardment mass spectrometry of sulphated oligosaccharides from ovine lutropin. Carbohyd Res. 209:33-50.

Griffin TJ, Seth G, Xie H, Bandhakavi S, Hu WS. (2007) Advancing mammalian cell culture engineering using genome-scale technologies. Trends Biotechnol. 25:401-408.

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

Hossler P, McDermott S, Racicot C, Chumsae C, Raharimampionona H, Zhou Y, Ouellette D, Matuck J, Correia I, Fann J, Li J. (2014) Cell culture media supplementation of uncommonly used sugars sucrose and tagatose for the targeted shifting of protein glycosylation profiles of recombinant protein therapeutics. Biotechnol Prog. 30:1419-1431.

Jayapal KP, Wlaschin KF, Hu WS, Yap MGS. (2007) Recombinant protein therapeutics from CHO cells - 20 years and counting. Chem Eng Prog. 103:40-47.

Kochanowski N, Blanchard F, Cacan R, Chirat F, Guedon E, Marc A, Goergen JL. (2008) Influence of intracellular nucleotide and nucleotide sugar contents on recombinant interferon-γ glycosylation during batch and fed-batch cultures of CHO cells. Biotechnol Bioeng. 100:721-733.

Krambeck FJ, Betenbaugh MJ. (2005) A Mathematical Model of N-Linked Glycosylation. Biotechnol Bioeng. 92:711-728.

Krambeck FJ, Bennum SV, Narang S, Choi S, Yarema KJ, Betenbaugh MJ. (2009) A mathematical model to derive N-glycan structures and cellular enzyme activities from mass spectrometric data. Glycobiology. 19:1163-1175.

Kremkow BG, Baik JY, MacDonald ML, Lee KH. (2015) CHOgenome.org 2.0: Genome resources and website updates. Biotechnol J. 10:931-938.

Kremkow BG, Lee KH. (2015) Sequencing technologies for animal cell culture research. Biotechnol Lett. 37:55-65.

Kremkow BG, Lee KH. (Submitted) Glyco-Mapper: A Chinese hamster ovary (CHO) genome-specific glycosylation prediction tool.

Kremkow B, Lee KH. (2013) Next-generation sequencing technologies and their potential impact on CHO cell-based biomanufacturing. Pharm Bioprocess. 1:455-465.

La Merie Business Intelligence. (2016) Blockbuster biologics 2015. R&D Pipeline News. 10:3-42.

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

Lis H, Sharon N. (1998) Lectins: carbohydrate-specific proteins that mediate cellular recognition. Chem Rev. 98:637-674.

Liu G, Neelamegham S. (2014) A Computational Framework for the Automated Construction of Glycosylation Reaction Networks. PLOS ONE. 9:e100939.

Malhotra R, Wormald MR, Rudd PM, Fischer PB, Dwer RA, Sim RB. (1995) Glycosylation changes of IgG associated with rheumatoid arthritis can activate complement via the mannose-binding protein. Nat Med. 1:237-243.

Mariño K, Bones J, Kattla JJ, Rudd PM. (2010) A systematic approach to protein glycosylation analysis: a path through the maze. Nat Chem Biol. 6:713-723.

Maszczak-Seneczko D, Sosicka P, Olczak T, Jakimowicz P, Majkowski M, Olczak M. (2013) UDP-N-acetylglucosamine Transporter (SLC35A3) Regulates Biosynthesis of Highly Branched N-glycans and Keratan Sulfate. J Biol Chem. 288:21850–21860.

Mechref Y, Muzikar J, Novotny MV. (2005) Comprehensive assessment of N-glycans derived from a murine monoclonal antibody: A case for multimethodological approach. Electrophoresis. 26:2034-2046.

Melmer M, Strangler T, Premstaller A, Lindner W. (2011) Comparison of hydrophilic-interaction, reversed-phase and porous graphitic carbon chromatography for glycan analysis. J Chromatogr A. 1218:118-123.

Nyberg GB, Balcarcel RR, Follstad BD, Stephanopoulos G, Wang DIC. (1998) Metabolic effects on recombinant interferon-y glycosylation in continuous culture of Chinese hamster ovary cells. Biotechnol Bioeng. 62:336-347.

O'Neill RA. (1996) Enzymatic release of oligosaccharides from glycoproteins for chromatographic and electrophoretic analysis. J Chromatogr A. 720:201-215.

Pharmaceutical Research and Manufacturers of America. (2015) 2015 biopharmaceutical research industry profile. PhRMA. 1-76.

Prater BD, Connelly HM, Qin Q, Cockrill SL. (2009) High-throughput immunoglobulin G N-glycan characterization using rapid resolution reverse-phase chromatography tandem mass spectrometry. Anal Biochem. 385:69-79.

Sandra K, Vandenheede I, Sandra P. (2014) Modern chromatographic and mass spectrometric techniques for protein biopharmaceutical characterization. J Chromatogr A. 1335:81-103.

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

Spahn PN, Hansen AH, Hansen HG, Arnsdorf J, Kildegaard HF, Lewis NE. (2016) A Markov chain model for N-linked protein glycosylation – towards a low parameter tool for model-driven glycoengineering. Metabol Eng. 33:52-66.

St. Amand MM, Radhakrishnan D, Robinson AS, Ogunnaike BA. (2014) Identification of manipulated variables for a glycosylation control strategy. Biotechnol Bioeng. 111:1957-1970.

Tateno H, Uchiyama N, Kuno A, Togayachi A, Sato T, Narimatsu H, Hirabayashi J. (2007) A novel strategy for mammalian cell surface glycome profiling using lectin microarray. Glycobiology. 17:1138-1146.

Umaña P, Bailey JE. (1997) A Mathematical Model of N-linked Glycoform Biosynthesis. Biotechnol Bioeng. 55:890-908.

US DHHS FDA. (2015) Quality considerations in demonstrating biosimilarity of a therapeutic protein product to a reference product. 1:1-19.

Vermassen T, Van Praet C, Vanderschaeghe D, Maenhout T, Lumen N, Callewaert N, Hoebeke P, Van Belle S, Rottey S, Delanghe J. (2014) Capillary electrophoresis of urinary prostate glycoproteins assists in the diagnosis of prostate cancer. Electrophoresis. 35:1017-1024.

Walsh G. (2007) Pharmaceutical biotechnology. John Wiley & Sons Inc. 1:1-11.

Walsh G. (2010) Biopharmaceutical benchmarks 2010. Nat Biotechnol. 28:917-924.

Walsh G, Jefferis R. (2006) Post-translational modifications in the context of therapeutic proteins. Nat Biotechnol. 24:1241-1252.

Wong NSC, Wati L, Nissom PM, Feng HT, Lee MM, Yap MGS. (2010) An investigation of intracellular glycosylation activities in CHO cells: Effects of nucleotide sugar precursor feeding. Biotechnol Bioeng. 107:321-326.

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

Yang Z, Wang S, Halim A, Schulz MA, Frodin F, Rahman SH, Vester-Christensen MB, Behrens C, Kristensen C, Vakhrushev SY, Bennett EP, Wandall HH, Clausen H. (2015) Engineered CHO cells for production of diverse, homogeneous glycoproteins. Nat Biotechnol. 33:842-844.

Zanetta JP, Pons A, Richet C, Huet G, Timmerman P, Leroy Y, Bohin A, Bohin JP, Trinel PA, Poulain D, Hofsteenge J. (2004) Quantitative gas chromatography/mass spectrometry determination of C-mannosylation of tryptophan residues in glycoproteins. Anal Biochem. 329:199-206.

Zhou X, Hong GY, Huang BB, Duan YM, Shen JY, Ni MW, Cong WT, Jin LT. (2014) Improved conditions for periodate/Schiff's base-based fluorescent staining of glycoproteins with dansylhydrazine in SDS-PAGE. Electrophoresis. 35:1439-1447.

Chapter 2

IMPACT OF SEQUENCING TECHNOLOGIES ON CHO CELL-BASED BIOPHARMACEUTICAL MANUFACTURING

2.1 Preface

This chapter is adapted from Kremkow and Lee (2013) with permission in accordance with the guidelines in the publishing contract. This chapter presents traditional Sanger sequencing as well as the next generation sequencing technologies 454 and Illumina. The applications of each technology as applied to the CHO cell-based biomanufacturing field are identified and summarized. The impact these technologies have had on CHO cell-based genomic, transcriptomic, proteomic, metabolomic, and multi-omic studies are analyzed and the applications towards industrial biomanufacturing are considered.

2.2 Abstract

There is growing interest in the possibility of harnessing detailed, mechanistic understanding of the biology of CHO cells to enhance the use of these cells in the manufacturing of biologics. Among the important questions about CHO cells are issues related to productivity, product quality attributes, and genomic stability. The advent of next generation DNA sequencing technologies provides an opportunity to characterize the genome of various host cells and to link genomic changes to phenotypes. In this chapter, I discuss some of the current and emerging technologies for genome sequencing, their initial applications to CHO bioprocessing, and provide context relative to other omic approaches.

2.3 Introduction

The 2012 recombinant biopharmaceutical market's global sales totaled more than $125 billion (La Merie 2013), from which mammalian cell line biopharmaceuticals were responsible for $86 billion and Chinese hamster ovary (CHO) cells are the host cell platform used to manufacture nearly 80% of the mammalian cell line-produced biopharmaceuticals. The biomanufacturing and cell line development community has made great strides to increase the capacity of CHO cells to produce recombinant proteins by more than 100-fold over the past 2-3 decades (Wurm 2004). These improvements are a result of engineering improvements throughout the entire bioprocess from cell line development to process understanding and purifications to control mechanisms (Wurm 2004; Butler 2005; Berlec and Strukelj 2013; Jenkins et al. 2009). In addition to enhancements in productivity, there

have been improvements to the ability to control product quality attributes (Jenkins et al. 2009; Zhu 2012). Nonetheless, it can be argued that there remain gaps in the CHO community's understanding of the detailed relationship between the cell culture conditions (raw materials, cell banking and expansion, growth in reactors, etc.) and stable, predictable outcomes in terms of both cell line productivity and product quality. Accordingly, there has been interest in the application of genomics, and related tools, to gain a better understanding of the biology happening inside CHO cells. The emergence of omic technologies over the past 20 years provides great opportunity for the field. However, the application of new and emerging technologies to the study of CHO cells for biomanufacturing invariably comes with a period of learning - to refine techniques, to understand benefits and drawbacks of various approaches, to develop new methods for data analysis - prior to substantial and transformative impact. Partnerships and collaboration among academic and industrial scientists will inevitably shorten the time needed to realize the benefits of the application of omics to biomanufacturing. In this chapter, I present an overview of the traditional and a couple next generation DNA sequencing technologies. In cases where the method has been applied to CHO cells, I also discuss the relevant applications.

2.4 Results and Discussion

2.4.1 Traditional DNA Sequencing

2.4.1.1 Traditional Sequencing Technology

The original "Sanger sequencing method" was developed in 1977 (Sanger et al. 1977) and some technological improvements have been made since then (Koutny et al.

2000; Emrich et al. 2002; Koster et al. 1996). The basic approach as used today relies on a DNA amplification based strategy with chain termination with dideoxynucleotides wherein individual molecules will terminate in a specific fluorescent molecule representing each of the nucleotide bases (A, T, G, or C). An electrophoretic separation of the molecule and subsequent detection, results in the ability to read a DNA sequence by reading the sequence of fluorophores. Because of limits with electrophoresis and with the chemistry of the reaction, one can only "read" a DNA sequence of a given length before the approach becomes unreliable. For modern versions of traditional sequencing, read lengths can reliably approach 900 base pairs (bp) of DNA. While 900 bp is much shorter than the typical gene or genome, it is long enough to provide the data needed to enable bioinformatic algorithms to assemble the data into gene sequences that are expressed or into whole genomes.

2.4.1.2 Application of Traditional Sequencing

The most common uses of traditional sequencing involve routine analyses that are performed as part of any molecular biology experiment. In the context of genomics, an important contribution of traditional sequencing technology is the ability to sequence genes that are expressed within an organism to facilitate analyses of gene expression. The collection of cDNAs from CHO cells and the subsequent sequencing of expressed sequence tags (ESTs) can allow one to identify possible CHO genes even without a complete CHO genome. This approach was successfully applied (Wlaschin et al. 2005) to CHO cells and the resulting EST sequences were searched (using the BLAST algorithm) against mouse, rat, and human genomes to demonstrate sequence

alignment between CHO ESTs and other mammalian genomes. In particular, it was observed that the strongest correlation existed between CHO and mouse sequences, resulting in the application of mouse microarrays to study CHO cells (Yee et al. 2008). Such work also ultimately led to the accumulation of enough CHO-specific cDNA sequence information to permit the design of CHO-specific microarrays.

Microarrays were used to explore the relationship between CHO gene expression and phenotypes of interest including high-productivity (Nissom et al. 2006), butyrate treatment effects (Yee et al. 2008; Kantardjieff et al. 2010), and the apoptosis pathways (Wong et al. 2006). For example, microarray analysis of a CHO-K1 suspension culture during lag, exponential, and stationary growth phases identified 1,400 mRNAs as differentially regulated at the stationary phase relative to the culture starting point (Bort et al. 2012). Further clustering analysis revealed gene groups with similar expression patterns (e.g. homologous recombination, Jak-STAT signaling pathway, spliceosome). In another study, microarray analysis of an IgG-producing CHO cell line at low temperature and butyrate conditions identified more than 900 differentially-expressed genes. Butyrate treatment was observed to increase protein production by inducing cell-cycle arrest, which coincides with the down-regulation of many cell cycle control genes. The altered culture conditions resulted in an increased IgG production rate, likely caused by an elevated cellular secretory capacity (Kantardjieff et al. 2010).

2.4.2 Next Generation DNA Sequencing

2.4.2.1 Next Generation Sequencing (NGS) Technologies

While traditional sequencing provided the technology platform needed to sequence and assemble the complete human genome in 2001 (Lander et al. 2001; Venter et al. 2001; Service 2006), as well as the mouse (Waterston et al. 2002) and rat (Gibbs et al. 2004) genomes, the cost of sequencing these genomes was in the billions and the timespan was more than a decade (Service 2006). Continued DNA sequencing method technology developments have substantially reduced the time required to collect sequence data while also reducing the cost. Motivated in part by a public effort to develop technologies to sequence a human genome for $1000 (Service 2006), next generation sequencing (NGS) technologies have emerged as alternatives to traditional sequencing: they include various new approaches to collecting genome-scale sequence information faster and for less cost than traditional sequencing; however, they also place an increased emphasis on bioinformatics to process and organize the resulting data. Nonetheless, NGS methods are being applied to study biological systems in several different ways. When used to sequence genomic libraries, NGS methods are a powerful method to resequence an organism, sequence organisms whose genome can be compared to a well-defined reference genome, or to collect data to perform a *de novo* genome assembly. They are also used to perform quantitative measures of changes in gene expression (e.g. mRNA expression) among a number of samples using an approach commonly referred to as RNASeq. Moreover, when these techniques are applied to small RNAs, they are used to catalog and quantify differences in expression in noncoding RNAs such as microRNAs.

NGS methods are commonly referred to by the name of the company associated with their development (even if that company was subsequently acquired and the technology rebranded) as shown in Table 2.1. The two most often used approaches (at least within the CHO community) are "454" (or Roche/454) and "Illumina" (or Solexa). It is beyond the scope of this chapter to describe in detail the nuances of these approaches and the reader is referred to recent articles that describe these methods in more detail (Rothberg and Leamon 2008; Bennett 2004; Mardis 2008; Ansorge 2009; Glenn 2011) as well as Appendix A for method characteristic comparisons. Nonetheless, here I provide a basic overview of 454 and Illumina before describing applications in biomanufacturing.

Table 2.1: SGS method statistics. Sequencing technology statistics demonstrate the 100- to 1,000-fold improvement of next generation sequencing compared to traditional Sanger sequencing. These technologies are commonly used for genomic and transcriptomic CHO studies.

Approach	**454**	**Illumina**	**Traditional**
Read Length	450-700 bp	36-250 bp	500-900 bp
Cost/Million Bases*	$10	$0.10	$2,400
Throughput/Run	700 Mb	47-600 Gb	0.9-87.0 kb
Run Time	23 hours	1.5-11 days	0.5-2 hours
Max. Reads/Run	1 million	1.5-6 billion	1, 16, 96
Advantage(s)	Moderate read length	High throughput, low cost	High quality, long read length
Drawback(s)	High relative cost for NGS, homopolymer sequence error rate, low throughput for NGS	Short read length	High cost, low throughput

* = Data adapted from (Liu et al. 2012; Quail et al. 2012)

454 is based upon pyrosequencing technology wherein the DNA sequence is determined by the emission of light that occurs when a complementary nucleotide is incorporated into a DNA molecule under certain conditions. By monitoring this process over hundreds of thousands of molecules in parallel, 454 takes advantage of multiplexing to accelerate the speed with which DNA sequence information is collected. The method has an average read length of 450-700 base pairs and can provide up to 1 million reads per run to yield approximately 450-700 Mb of data (Rothberg and Leamon 2008) per run. Among the advantages of this approach are that

the technology provides high quality data, good genome coverage, and a lower cost (per base pair of sequence information) than traditional sequencing. However, the method yields shorter read lengths than traditional sequencing which means that it is important to employ relevant bioinformatic algorithms to assemble DNA sequence information together. Other issues with this approach are that homopolymers are difficult to sequence, and there is a decreased throughput of data collection compared to some other NGS methods.

Illumina technology is based upon "sequencing by synthesis" and also takes advantage of parallel processing of DNA sequence strands to generate sequence information. Strands of DNA (from a genomic library) are attached to a surface and nucleotides are added. After each nucleotide addition, images are captured of the location of specific nucleotide addition based on fluorescence. A series of images can be analyzed to determine the DNA sequence of billions of strands of DNA in a single instrument run. Among the advantages of this approach are the very large amounts of data that can be collected (up to 300 or 600 Gb of data), relatively low error rates, and the very low cost of sequencing (on a per base pair of sequence information basis). Among the limitations of the current form of the technology are the relatively smaller read lengths (below 250 bp) which makes sequencing through repeat regions, and assembly of information, even more challenging than other NGS methods (Ansorge 2009). A number of bioinformatic tools have been developed to address some of the challenges posed by the shorter NGS reads (Li et al. 2009; Simpson et al 2009; Trapnell and Salzberg 2009; Flicek and Birney 2009; Miller et al. 2010).

2.4.2.2 Applications of NGS

NGS approaches have been applied to biomanufacturing-related questions for a number of years. However, one could argue that the most important advance to the field has been the use of NGS to sequence CHO genomes (Hammond et al. 2011; Xu et al. 2011; Lewis et al. 2013) as well as the Chinese hamster (Lewis et al. 2013) and its chromosomes (Brinkrolf et al. 2013). Because NGS approaches are effective not only at sequencing genomes, but also at monitoring changes in expression profiles of mRNA and microRNA, their application to problems relevant to biomanufacturing began before the CHO and Chinese hamster genomes were made available. A list of CHO-based NGS studies I have compiled is provided in Table 2.2. For example, there were significant efforts directed at the use of 454 to generate as many genomic reads from CHO as possible to collect and expand knowledge regarding CHO ESTs for the development of tools to facilitate transcriptome analysis. 400,000 reads from CHO with an average length of 212 bp (Kantardjieff et al. 2009) were aligned with more than 34,000 available ESTs to extend the sequence of 70% of the ESTs an average of 150 bp.

Table 2.2: CHO-based next-generation sequencing publications. Table of NGS-based CHO studies including type of study, cell lines, technology platform, and focus area. Proteomics studies reliant upon the NGS-based CHO genome are included.

Omics	Cell Line(s)	NGS Technology Used	Sequences	Reference
G	CHO: SEAP	Illumina	Genes	Hammond et al. 2011
G/T	CHO-K1	Illumina	Genome, glycosylation and viral susceptability pathway genes	Xu et al. 2011
G/T	Chinese hamster, CHO-K1, DG44, CHO-S	Illumina	Genomes, apoptosis pathway genes	Lewis et al. 2013
G	Chinese hamster	Illumina	Chromosomes (genome)	Brinkrolf et al. 2013
T	Parental CHO	454	ESTs	Kantardjieff et al. 2009
T/G	CHO: IgG	Illumina	Transcriptome, butyrate-affected pathway genes	Birzele et al. 2010
T	CHO: IgG	Illumina	Transcriptome	Jacob et al. 2010
T	CHO-K1	454	Transcriptome, N-glycosylation pathway genes, and splice variants	Becker et al. 2011
T	CHO-K1, DG-44	Illumina	miRNA	Johnson et al. 2011
T	CHO-K1, CHO-DUXB11	Illumina	miRNA transcriptome: novel and conserved	Hackl et al. 2011
T	CHO-K1	Illumina	miRNA genomic loci and precursor miRNA	Hackl et al. 2012
T	CHO-K1, CHO: SEAP, CHO: tPA	Illumina	miRNA	Hammond et al. 2012
T	CHO-K1, CHO-DUKXB-11	Illumina	piRNAs and piRNA clusters	Gerstl et al. 2013
P	CHO-K1: SEAP	CHO gene database	Improved proteome identification based on genome sequences	Meleady et al. 2012
P	CHO-K1	CHO gene database	Codon frequency, gene ontology, post translational modifications	Baycin-Hizal et al. 2012

(G=Genomics, T=Transcriptomics, P=Proteomics)

An IgG-producing CHO cell line grown under butyrate conditions was studied with Illumina to identify genes responsible for enhanced productivity (Birzele et al. 2010). More than 13,000 CHO genes were sequenced and annotated, using genomic information from similar organisms, and approximately 5,000 novel CHO genes were identified and added to their CHO model. In this same study, the transcriptome was analyzed for gene clusters affected by butyrate treatment and suggested that the down-regulated genes were related to cell cycle check point control, mitotic check point control, and the initiation and elongation phases of DNA replication processes (Birzele et al. 2010). While these observations demonstrated agreement with other butyrate transcriptomic analyses in literature (Yee et al. 2008; Kantardjieff et al. 2010; Gatti et al. 2007), the additional knowledge gained may facilitate a better understanding of high productivity phenotypes for cell line development.

Illumina analysis of a secreted alkaline phosphatase-producing line yielded 3.57 million contigs and provided CHO-specific sequence information for 18,000-19,000 metabolic process, cellular signaling, and transport orthologs (Miller et al. 2010). This approach identified nearly 5,000 additional CHO genes without a reference CHO genome and demonstrated the possibility of using NGS to sequence an entire CHO cell line genome.

Transcriptomic analysis of an IgG-producing CHO line yielded 55 million sequencing reads that were mapped to an existing CHO EST-derived unigene set and several public sequence databases (Jacob et al. 2010). The transcript abundance varied up to six orders of magnitude, while the coverage across the transcript lengths varied to a far lesser extent (Jacob et al. 2010). While the sequencing was successful,

methods for coefficient of variation reduction related to the use of NGS results for transcript measurements and gene expression were addressed, but not fully resolved.

Transcriptomes from multiple recombinant CHO cell lines under various cultivation conditions were investigated with 454 technology (Becker et al. 2011). The findings reinforced the idea that there is some reasonable amount of CHO gene sequence similarity to mouse sequences. The gene transcript levels relevant to the central carbohydrate metabolism and glycosylation pathways were measured, which enabled construction of accurate model pathways. For each section of the N-glycosylation pathway, at least one gene was measured. The set of assembled and annotated CHO cell genes purposed for CHO cell line transcript analysis was made available and 70% of the sequences are most similar to the mouse transcriptome relative to human and rat. Approximately 6,700 genes were covered by the mouse sequences by at least 95% of their sequence length (Becker et al. 2011), confirming the work of Wlaschin (Wlaschin et al. 2005).

MicroRNA (miRNA) is one class of noncoding RNAs (ncRNAs) that has a unique and characteristic secondary structure (Bartel 2009). Precursor miRNA transcripts are nearly 70 nucleotides and the processed, mature miRNA sequences are typically between 20 and 24 nucleotides (Johnson et al. 2011) in length, which makes them well-suited for analysis using the Illumina platform. Mature miRNAs control the fate of gene expression via post-transcriptional repression of mRNA translation or destabilization (Huntzinger and Izaurralde 2011) and as a result, they have great potential in cell characterization and engineering applications. For example, miRNA sequences that are expressed during cultivation may influence a range of cellular processes that control productivity, product quality attributes, and growth. Many

miRNA sequences are conserved across species and others may be unique to species. One of the ongoing issues in the life science community is the ability to catalog and understand each of the known miRNAs. In the past few years, the number of known CHO miRNA sequences has significantly increased. Expressed miRNA can be extracted from CHO cell lines and sequenced. The resulting sequences can be compared against general databases of known miRNAs (e.g. miRBase) to confirm sequences already known or conserved across species. By linking the sequence to the experimental conditions used that resulted in the miRNA expression, one can begin to establish a link between miRNA and phenotype. The number of CHO miRNA sequences has gone from 260 (Kantardjieff et al. 2009) to 350 (Johnson et al. 2011) to 387 (Hackl et al. 2011) within the past few years alone. Of the nearly 400 known CHO miRNA sequences, 350 of which are conserved and 235 of which have a specified function (Hackl et al. 2011).

In addition to experimental approaches, bioinformatic algorithms can also be used to predict miRNA sequences from an established genome. With the sequencing of the CHO-K1 genome (Xu et al. 2011), the ability to computationally predict CHO miRNAs became possible and resulted in 415 miRNA sequence identifications (Hackl et al. 2012). The locations of 365 structures were cataloged, 319 of which are expressed, mature CHO miRNAs that were verified and assigned to miRBase (Hackl et al. 2012). Relative genomic locations have also been used in the CHO-K1 genome in an attempt to discover additional miRNA sequences based upon identified miRNA scaffold organization (Hammond et al. 2012). Post CHO-K1 genome availability, bioinformatics tools identified 190 miRNAs as conserved CHO miRNAs between the CHO-K1, mouse, rat, and human genomes, of which more than 80% exhibited

differential expression across two recombinant CHO cell lines (Hammond et al. 2012). Moving forward, NGS approaches offer an unprecedented ability to interrogate changes in both mRNA and miRNA expression that are related to phenotypes of interest (Bort et al. 2012).

Unlike miRNA, PIWI (a class of proteins) interacting RNAs (piRNAs) are a poorly understood class of small ncRNAs, which likely mediate RNA silencing and repress transposable elements, protecting the genome's integrity (Thomson and Lin 2009; Gerstl et al. 2013). piRNA function may affect the prolonged stability of genetically modified CHO cell lines, in addition to cellular processes and metabolic pathways. Computational analysis of small RNA sequencing data predicted 540 piRNA clusters, consisting of nearly 26,000 piRNA sequences (Gerstl et al. 2013). piRNA sequence expression was measured across six CHO cell lines, including adherent, suspension adapted, and recombinant CHO-K1 and CHO-DUKXB11 cell lines, using the published CHO-K1 genome as a reference (Gerstl et al. 2013). This initial analysis of CHO piRNA indicated the potential of piRNAs as tools for cell line development and genetic engineering.

As mentioned earlier, perhaps the most important contribution of NGS to the CHO biomanufacturing community has been the establishment of the CHO-K1 and Chinese hamster genomes (Xu et al. 2011; Lewis et al. 2013; Brinkrolf et al. 2013). The draft CHO-K1 genomic sequence consisted of 2.45 Gb and was assembled into 24,383 genes (Xu et al. 2011). This catalog of genes permitted genetic modification of CHO-specific target DNA sequences. The CHO-K1 genes were analyzed by comparative genomic analysis with the human, mouse, and rat genomes, which confirmed that the mouse genome demonstrated the greatest similarity (Xu et al.

2011). The draft genome enabled immediate analysis of genes relevant to CHO biomanufacturing such as those related to product quality attributes such as glycosylation. In that study (Xu et al. 2011), the number of CHO genes that were homologous to human glycosylation genes suggested that CHO cells have the potential to perform 99% of the glycosylation reactions that humans perform. However, transcriptome analysis of CHO cells further suggested that only about half of the CHO glycosylation genes were actually expressed under any conditions, meaning that the other half may be silenced. Such analyses based on a draft genome promise a more detailed, molecular understanding, of the behavior of CHO cells leading to improvements in bioprocessing in the future. However, there are also a number of unaddressed challenges that emerged with the CHO genome, partly as a result of the use of NGS methods.

One important issue stemming from the current CHO-K1 (Xu et al. 2011) and Chinese hamster genomes (Lewis et al. 2013) is that the assembled DNA sequences have not been aligned onto chromosomes. An important consideration in the CHO genome is the lack of chromosomal stability that has been observed (Wurm and Hacker 2011). Moreover, there is significant genomic drift (Wurm and Hacker 2011). Indeed, these are properties that the biomanufacturing community has embraced in the application of CHO cells because of the ability to reasonably quickly adapt CHO cells to various growth conditions. However, once established, it would be advantageous to have host cells that have minimal chromosomal or genomic changes. To facilitate analysis of chromosomal or genomic rearrangements, it is important for the CHO community to have a physical mapping of a reference genome. One way to help build

a physical map of the CHO genome is to use a bacterial artificial chromosome (BAC) library (Omasa et al. 2009).

An initial CHO BAC-based map identified twenty different chromosomes and high aneuploidy was observed. This library was used to obtain a detailed physical chromosomal map of the CHO-DG44 cell line utilizing fluorescence *in situ* hybridization imaging of the randomly selected BAC clones. For eight of the twenty chromosomes identified, chromosomal rearrangements did not occur between CHO-DG44, CHO-K1, and Chinese hamster lung cells. The conservation without large rearrangement suggests their genetic importance and resultant stability (Cao et al. 2012); however, it was not possible to identify what genes were located on these eight chromosomes. The recent publication of sequences for individual Chinese hamster chromosomes (Brinkrolf et al. 2013) provides a critical step forward for the community because it includes the sequences of each of the Chinese hamster chromosomes independent of the others.

A second important issue from the CHO-K1 genome (Xu et al. 2011) is the need for ongoing updates to the assembly and annotation. While the initial draft genome is assembled and annotated with the aid of humans, the majority of the work is done by bioinformatic algorithms. The genome communities associated with humans and other organisms have developed mechanisms to make updates and corrections to their genomes and the CHO genome community has recently established a framework (Hammond et al. 2012) to facilitate similar efforts. Given the diversity of cell lines and the known issues with significant genomic and chromosomal variability among CHO cells, the identification and establishment of a definitive reference genome is essential for the community.

To better understand the genomic diversity and help establish a reference, the draft genomes for the Chinese hamster (*Cricetulus griseus),* the CHO-DG44, CHO-S, and three other CHO cell lines were sequenced and the CHO-K1 cell line was resequenced (Lewis et al. 2013). Annotation of all cell lines and nucleotide-resolution analysis of the CHO cell line genotypic differences was completed. Comparative genomics identified copy number variations and 3.7 million single-nucleotide polymorphisms (SNPs) between the different cell lines, many of which affected genes relevant to bioprocessing pathways, such as apoptosis (Lewis et al. 2013). In an attempt to determine the genomic structure, the sequences were aligned to published BACs and filtered; however, only 26% of the genomic sequence was reliably localized to specific hamster chromosomes (Lewis et al. 2013).

Following the release of the CHO-K1 genome, an international academic and industrial collaboration developed CHOgenome.org to facilitate accessibility of the genomic data and the development of genomic tools for the *Cricetulus griseus* and CHO cell communities (Hammond et al. 2012). The current list of tools offered includes BLAST searches, individual gene searches, and visual representation of the CHO-K1 genome assemblies (Hammond et al. 2012), along with a CHO proteome database. However, the sequencing of a number of CHO-related genomes creates challenges in terms of comparative genomics. That is, there is a need for tools to facilitate the analysis of multiple genomes by a given user. For example, it may be desirable to compare the genome of a host cell early in culture versus late in culture or to compare the genome of a proprietary host cell with that of CHO-K1 and of the Chinese hamster. Tools that facilitate analysis of events as large as chromosomal rearrangements and as small as SNPs would provide users the opportunity to link

genome information to observed phenotypes. However, such tools have not yet been created or adapted for CHO-specific applications.

2.4.3 Emerging Sequencing Approaches

The pace of DNA sequencing technology development has continued and there is now a new generation of technologies available. These approaches, which offer new and greater amounts of data for similar or less cost per run, are designed with single-molecule or electrochemical platforms, compared to the NGS platforms that use fluorescent signals and PCR amplification. Among the new methods (Korlach et al. 2010; Rhee and Burns 2006; Merriman et al. 2012; Liu et al. 2012; Quail et al. 2012; Roberts et al. 2013; McCarthy 2010; Clarke et al. 2009) are those developed by Pacific Biosciences (PacBio), Life Technologies (Ion Torrent), and Oxford Nanopore Technologies (nanopore). The PacBio approach involves single molecule sequencing based on an immobilized polymerase in a cell designed for single molecule, real time (SMRT) detection, which has potential applications in the bioprocessing field include studies of epigenetic regulation of heterologous gene regulation and genome-wide structural variation (McCarthy 2010). Ion Torrent is a semiconductor platform based sequencing approach that relies on a pH probe for detection and takes advantage of the fact that hydrogen ions are released as nucleotides are incorporated into a growing DNA chain (Merriman et al. 2012). The nanopore sequencing approach also relies on an immobilized enzyme, staphylococcal α-hemolysin, as the nanopore through which a DNA molecule is sequenced by passing through the pore and across an electric potential field (Rhee and Burns 2006). While none of the emerging and third generation sequencing technologies have yet been applied to CHO studies relevant to

the biomanufacturing community (at the time of this publication), these approaches will certainly see widespread application in the near future. Further details regarding these techniques are included in Appendix A.

2.4.4 Other Omics

While NGS has had the most dramatic and obvious impact on the CHO community in the past few years, there are a number of important and parallel omics approaches that are also necessary for a complete understanding of how CHO biology is linked to enhanced productivity and product quality attributes. This fact is validated by the significant amount of literature that has been published regarding this topic (Ahn and Antoniewicz 2012; Chong et al. 2010; Luo et al. 2012; Hayduk and Lee 2005; Pascoe et al. 2007; Baik et al. 2006; Crea et al. 2006; North et al. 2010; Tateno et al. 2007). The important issue to consider is that genome sequencing and genomics alone may not provide enough information about observed phenotypes. Certainly the genes and other features (e.g. miRNA) that are expressed significantly influence cell behaviors. The mRNAs lead to protein expression and proteins have diverse functions including structural roles, metabolism, and many other cell processes. Ultimately, a true understanding of the basis for a given phenotype may rely not only on the ability to capture the genome of the cell line at that moment in time, but also on transcript analysis, proteomics, metabolomics, and other measures. The CHO community has applied many of these other techniques individually to understand phenotypes, but very few studies to date have integrated data from NGS studies together with other omic methods - efforts which rely heavily on bioinformatics because of the large volume of data created by NGS. Clarke et al. (Clarke et al. 2012) is one example to the

contrary where transcriptomics (using NGS methods) and proteomics were both used to study the CHO cell growth rate.

In one metabolomics study, an *in silico* model was used with metabolomic analysis to understand CHO intracellular fed-batch culture mechanisms. The identified, growth limitation metabolites were associated with the glutathione, glycerophospholipid, and energy pathways (Selvarasu et al. 2012). The *in silico* model was used to obtain a greater understanding of these affected pathway fluxes, the results of which were in good agreement with aging culture glycolysis and TCA cycle flux details, resulting in the identification of novel, growth-related mechanisms. The *in silico* model used was not originally developed from the annotated CHO genome, but rather from the mouse genome (Sheikh et al. 2005), refined and validated with mouse hybridoma cell observations (Selvarasu et al. 2010), and expanded for CHO with annotated CHO cDNA, as the CHO genome and gene function identification was ongoing. A completely CHO genome-based model would potentially enhance these results, future metabolomics studies, and lead to new CHO culture improvements.

Proteomic analysis has also been applied to the study of biopharmaceutical production cell lines for many years but only recently has it been combined with NGS datasets. Proteomics studies typically involve the use of mass spectrometry to link changes in observed proteins to their underlying genes. The availability of a sequenced CHO genome, which was facilitated by NGS, has provided a means to improve the efficiency by which mass spectra are assigned to gene sequences (Meleady et al. 2012). For example, two CHO specific databases were used for CHO protein identification, including the CHO-K1 genome database. Identification using this database increased the number of identified proteins by 35%, which further increased

to 47% with the addition of a second CHO specific database (Meleady et al. 2012). In another recent analysis of the CHO proteome based on the CHO K1 genome, the proteome, secretome, and glycoproteome contained 6,164 grouped proteins (Baycin-Hizal et al. 2012), an eight-fold increase in the number of CHO proteins identified. This increase was attributable to both an improved cell lysate fractionation method as well as the use of an organism-specific sequence database. More importantly, the availability of a detailed proteome dataset permits a better understanding of codon frequency in CHO, the degree of pathway enrichment, and of possible post-translational modifications. For example, codon frequency in CHO was observed to be distinct from humans (Baycin-Hizal et al. 2012). The degree of pathway enrichment was obtained from combined proteomic and transcriptomic (mRNA) data sets, highlighted by the enrichment of the protein processing and apoptosis pathways and depletion of the steroid hormone and glycosphingolipid metabolism pathways. The cataloged post-translational modifications included 504 N-acetylation proteins and 1,292 N-glycosylated proteins.

2.5 Concluding Remarks

The CHO biomanufacturing community is at the beginning of the genomics era. The unprecedented ease with which one can collect DNA sequence information will enable a deeper understanding of the relationship between the genome and phenotypes. However, the pace with which data is generated is increasing and there are bottlenecks in the ability to analyze the data. As a result, there may be an increasing emphasis on bioinformaticians who can assist in the interpretation and understanding of these large datasets. Moreover, there is an urgent need for a well-

defined reference genome that is stable and that the community can use as a foundation for genomics-based studies. Once established, individual teams can employ methods to study the genome, epigenome, transcriptome, proteome, and metabolome as part of their efforts to understand cellular phenotypes. However, making biological inferences that will lead to targets for cellular engineering to modify cell productivity, product quality attributes, and the stability of cell lines in a predictable manner may take many years. Cooperation and collaboration among the academic and industrial scientific community will be essential for the CHO community to fully realize the potential of the genomics era for the production of biologics.

The CHO biomanufacturing community is at the start of a new era. The availability of the first drafts of a number of relevant genomes and the low cost of sequencing various host cells provides a basic foundation for the community to better understand the molecular basis for issues related to productivity, product quality, and stability (of productivity, of product quality, and of viability). However, many challenges are also emerging. First, the community does not yet have a stable, well-defined and characterized reference genome. The recent sequencing of the Chinese hamster and its chromosomes provides the basis for this moving forward, but the need to correct annotations remains an ongoing challenge even in the human genome community. Second, there are relatively few tools available to compare genomes. For example, tools to easily compare the genome of a proprietary host cell versus CHO K1 versus the Chinese hamster, or to compare the genome of a host cell early in culture versus late in culture, do not exist. Third, having a genome is most useful when placed in the context of other omic data for a given cell (transcriptomics, proteomics, fluxomics, etc.) and there are not yet tools available to simply integrate information

across these datasets. Despite these and other challenges, NGS has helped move the CHO biomanufacturing community forward towards a time when host cells can make any given product and can be reliably and predictably customized and designed to ensure high productivity of specific product quality attributes that are stably expressed by cells.

2.6 Acknowledgements

I would like to thank the National Science Foundation (1247394), the National Institutes of Standards and Technologies (60NANB11D185), and the International CHO Genome Community for financial support.

REFERENCES

Ahn WS, Antoniewicz MR. (2012) Towards dynamic metabolic flux analysis in CHO cell cultures. Biotechnol J. 7:61-74.

Ansorge WJ. (2009) Next-generation DNA sequencing techniques. New Biotechnol. 25:195-203.

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

Bartel DP. (2009) MicroRNAs: target recognition and regulatory functions. Cell. 136:215-233.

Baycin-Hizal D, Tabb DL, Chaerkady R, Chen L, Lewis NL, Nagarajan H, Sarkaria V, Kumar A, Wolozny D, Colao J, Jacobson E, Tian Y, O'Meally RN, Krag SS, Cole RN, Palsson BØ, Zhang H, Betenbaugh M. (2012) Proteomic analysis of Chinese hamster ovary cells. J Proteome Res. 11:5265-5276.

Becker J, Hackl M, Rupp O, Jakobi T, Schneider J, Szczepanowski R, Bekel T, Borth N, Goesmann A, Grillari J, Kaltschmidt C, Noll T, Pühler A, Tauch A, Brinkrolf K. (2011) Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. J Biotechnol. 156:227-235.

Bennett S. (2004) Solexa ltd. Pharmacogenomics. 5:433-438.

Berlec A, Strukelj B. (2013) Current state and recent advances in biopharmaceutical production in Escherichia coli, yeasts and mammalian cells. J Ind Microbiol Biot. 40:257-274.

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

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

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

Butler M. (2005) Animal cell cultures: recent achievements and perspectives in the production of biopharmaceuticals. Appl Microbiol Biot. 68:283-291.

Cao Y, Kimura S, Itoi T, Honda K, Ohtake H, Omasa T. (2012) Construction of BAC-based physical map and analysis of chromosome rearrangement in Chinese hamster ovary cell lines. Biotechnol Bioeng. 109:1357-1367.

Chong WPK, Reddy SG, Yusufi FNK, Lee DY, Wong NSC, Heng CK, Yap MGS, Ho YS. (2010) Metabolomics-driven approach for the improvement of Chinese hamster ovary cell growth: overexpression of malate dehydrogenase II. J Biotechnol. 147:116-121.

Clarke C, Henry M, Doolan P, Kelly S, Aherne S, Sanchez N, Kelly P, Kinsella P, Breen L, Madden SF, Zhang L, Leonard M, Clynes M, Meleady P, Barron N. (2012) Integrated miRNA, mRNA and protein expression analysis reveals the role of post-transcriptional regulation in controlling CHO cell growth rate. BMC Genomics, 13:656.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 4:265-270.

Crea F, Sarti D, Falciani F, Al-Rubeai M. (2006) Over-expression of hTERT in CHOK1 results in decreased apoptosis and reduced serum dependency. J Biotechnol. 121:109-123.

Davies SL, Lovelady CS, Grainger RK, Racher AJ, Young RJ, James DC. (2013) Functional heterogeneity and heritability in CHO cell populations. Biotechnol Bioeng. 110:260-274.

Emrich CA, Tian HJ, Medintz IL, Mathies RA. (2002) Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. Anal Chem. 74:5076-5083.

Flicek P, Birney E. (2009) Sense from sequence reads: methods for alignment and assembly. Nat Methods. 6:S6-S12.

Gatti MD, Wlaschin KF, Nissom PM, Yap M, Hu WS. (2007) Comparative transcriptional analysis of mouse hybridoma and recombinant Chinese hamster ovary cells undergoing butyrate treatment. J Biosci Bioeng. 103:82-91.

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

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera, Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Venter JC, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtel K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bouffard GG, NHGRI, de Jong PJ, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 428:493-521.

Glenn TC. (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour. 11:759-769.

Hackl M, Jadhav V, Jakobi T, Rupp O, Brinkrolf K, Goesmann A, Pühler A, Noll T, Borth N, Grillari J. (2012) Computational identification of microRNA gene loci and precursor microRNA sequences in CHO cell lines. J Biotechnol. 158:151-155.

Hackl M, Jakobi T, Blom J, Doppmeier D, Brinkrolf K, Szczepanowski R, Bernhart S, Siederdissen CH, Bort JAH, Wieser M, Kunert R, Jeffs S, Hofacker IL, Goesmann A, Pühler A, Borth N, Grillari J. (2011) Next-generation sequencing of the Chinese hamster ovary microRNA transcriptome: identification, annotation and profiling of microRNAs as targets for cellular engineering. J Biotechnol. 153:62-75.

Hammond S, Kaplarevic M, Borth N, Betenbaugh MJ, Lee KH. (2012) Chinese hamster genome database: an online resource for the CHO community at www.CHOgenome.org. Biotechnol Bioeng. 109:1353-1356.

Hammond S, Swanberg JC, Kaplarevic M, Lee KH. (2011) Genomic sequencing and analysis of a Chinese hamster ovary cell line using Illumina sequencing technology. BMC Genomics. 12:67.

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

Hayduk EJ, Lee KH. (2005) Cytochalasin D can improve heterologous protein productivity in adherent Chinese hamster ovary cells. Biotechnol Bioeng. 90:354-364.

Huntzinger E, Izaurralde E. (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet. 12:99-110.

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

Jenkins N, Meleady P, Tyther R, Murphy L. (2009) Strategies for analysing and improving the expression and quality of recombinant proteins made in mammalian cells. Biotechnol Appl Bioc. 53:73-83.

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

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

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

Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW. (2010) Real-time DNA sequencing from single polymerase molecules. Method Enzymol. 472:431-455.

Koster H, Tang K, Fu DJ *et al.* (1996) A strategy for rapid and efficient DNA sequencing by mass spectrometry. Nat Biotechnol. 14:1123-1128.

Koutny L, Schmalzing D, Salas-Solano O, El-Difrawy S, Adourian A, Buonocore S, Abbey K, McEwan P, Matsudaira P, Ehrlich D. (2000) Eight hundred base sequencing in a microfabricated electrophoretic device. Anal Chem. 72:3388-3391.

La Merie Business Intelligence. (2013) Blockbuster biologics 2012. R&D Pipeline News. 7:2-28.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond, C, Rosetti M, Santos R, Sheridan A, Sougnez C, Strange-Thomann N, Stojanovic N, Subramanian A, Wyman D, *et al.* (2001) Initial sequencing and analysis of the human genome. Nature. 409:860-921.

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

Li RQ, Li YR, Kristiansen K, Wang J. (2008) SOAP: short oligonucleotide alignment program. Bioinformatics. 24:713-714.

Liu L, Li YH, Li SL, Hu N, He Y, Pong R, Lin D, Lu L, Law M. (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012:1-11.

Luo J, Vijayasankaran N, Autsen J, Santuray R, Hudson T, Amanullah A, Li F. (2012) Comparative metabolite analysis to understand lactate metabolism shift in Chinese hamster ovary cell culture process. Biotechnol Bioeng. 109:146-156.

Mardis ER. (2008) Next-generation DNA sequencing methods. Annu Rev Genom Hum G. 9:387-402.

McCarthy A. (2010) Third generation DNA sequencing: Pacific Biosciences' single molecule real time technology. Chem Biol. 17:675-676.

Meleady P, Hoffrogge R, Henry M, Rupp O, Bort JH, Clarke C, Brinkrolf K, Kelly S, Müller B, Doolan P, Hackl M, Beckmann TF, Noll T, Grillari J, Barron N, Pühler A, Clynes M, Borth N. (2012) Utilization and evaluation of CHO-specific sequence databases for mass spectrometry based proteomics. Biotechnol Bioeng. 109:1386-1394.

Merriman B, Rothberg JM, Ion Torrent R, Team D. (2012) Progress in Ion Torrent semiconductor chip based sequencing. Electrophoresis. 33:3397-3417.

Miller JR, Koren S, Sutton G. (2010) Assembly algorithms for next-generation sequencing data. Genomics. 95:315-327.

Nissom PM, Sanny A, Kok YJ, Hiang YT, Chuah SH, Shing TK, Lee YY, Wong KTK, Hu WS, Yap MGS, Philp R. (2006) Transcriptome and proteome profiling to understanding the biology of high productivity CHO cells. Mol Biotechnol. 34:125-140.

North SJ, Huang HH, Sundaram S, Jang-Lee J, Etienne T, Trollope A, Chalabi S, Dell A, Stanley P, Haslam SM. (2010) Glycomics profiling of Chinese hamster ovary cell glycosylation mutants reveals N-glycans of a novel size and complexity. J Biol Chem. 285:5759-5775.

Omasa T, Cao YH, Park JY, Takagi Y, Kimura S, Yano h, Honda K, Asakawa S, Shimizu N, Ohtake H. (2009) Bacterial artificial chromosome library for genome-wide analysis of Chinese hamster ovary cells. Biotechnol Bioeng. 104:986-994.

Pascoe DE, Arnott D, Papoutsakis ET, Miller WM, Andersen DC. (2007) Proteome analysis of anti body-producing CHO cell lines with different metabolic profiles. Biotechnol Bioeng. 98:391-410.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 13:341.

Rhee M, Burns MA. (2006) Nanopore sequencing technology: research trends and applications. Trends Biotechnol. 24:580-586.

Roberts RJ, Carneiro MO, Schatz MC. (2013) The advantages of SMRT sequencing. Genome Biol. 14:405.

Rothberg JM, Leamon JH. (2008) The development and impact of 454 sequencing. Nat Biotechnol. 26:1117-1124.

Sanger F, Nicklen S, Coulson AR. (1977) DNA sequencing with chain-terminating inhibitors. P Natl Acad Sci USA. 74:5463-5467.

Selvarasu S, Ho YS, Chong WPK, Wong NSC, Yusufi FNK, Lee YY, Yap MGS, Lee DY. (2012) Combined *in silico* modeling and metabolomics analysis to characterize fed-batch CHO cell culture. Biotechnol Bioeng. 109:1415-1429.

Selvarasu S, Karimi IA, Ghim GH, Lee DY. (2010) Genome-scale modeling and *in silico* analysis of mouse cell metabolic network. Mol Biosyst. 6:152-161.

Service RF. (2006) Gene sequencing - The race for the $1000 genome. Science. 311:1544-1546.

Sheikh K, Forster J, Nielsen LK. (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus*. Biotechnol Progr. 21:112-121.

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

Tateno H, Uchiyama N, Kuno A, Togayachi A, Sato T, Narimatsu H, Hirabayashi J. (2007) A novel strategy for mammalian cell surface glycome profiling using lectin microarray. Glycobiology. 17:1138-1146.

Thomson T, Lin HF. (2009) The Biogenesis and function of PIWI proteins and piRNAs: progress and prospect. Annu Rev Cell and Dev Bi. 25:355-376.

Trapnell C, Salzberg SL. (2009) How to map billions of short reads onto genomes. Nat Biotechnol. 27:455-457.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson D, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Miklos GLG, Nelson C, Broder S, *et al.* (2001) The sequence of the human genome. Science. 291:1304-1351.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520-562.

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

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

Wurm FM, Hacker D. (2011) First CHO genome. Nat Biotechnol. 29:718-720.

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

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

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

Yee JC, Wlaschin KF, Chuah SH, Nissom PM, Hu WS. (2008) Quality assessment of cross-species hybridization of CHO transcriptome on a mouse DNA oligo microarray. Biotechnol Bioeng. 101:1359-1365.

Zhu J. (2012) Mammalian cell protein expression for biopharmaceutical production. Biotechnol Adv. 30:1158-1170.

Chapter 3

IMPROVING CHOGENOME.ORG TO BE THE CENTRAL CHO GENOME RESOURCE

3.1 Preface

This chapter is adapted from Kremkow et al. (2015) with permission (see Appendix D). In this chapter, I present the *Cricetulus griseus* information added to CHOgenome.org since its release as well as additional tools and the upgraded website. The recorded website usage is analyzed and published studies using CHOgneome.org are detailed, demonstrating the impact CHOgenome.org has had upon the CHO community. The genome viewer upgrade to JBrowse was performed by Madolyn MacDonald-Stinner and database coding was performed by Eric Garrison.

3.2 Abstract

Chinese hamster ovary (CHO) cells are a major host cell line for the production of therapeutic proteins, and the CHO-K1 cell and Chinese hamster (CH) genomes have recently been sequenced using next-generation sequencing methods. CHOgenome.org was launched in 2011 (version 1.0) to serve as a database repository and to provide bioinformatics tools for the CHO community. CHOgenome.org (version 1.0) maintained GenBank CHO-K1 genome data, identified CHO -omics literature, and provided a CHO-specific BLAST service. I have implemented recent major updates to CHOgenome.org (version 2.0) including new sequence and annotation databases for both CHO and CH genomes, a more user-friendly website, and new research tools, specifically a proteome browser and a genome viewer. CHO cell line specific sequences and annotations facilitate cell line development opportunities, several of which are discussed. Moving forward, CHOgenome.org will host the increasing amount of CHO -omics data and continue to make useful bioinformatics tools available to the CHO community.

3.3 Introduction

Chinese hamster ovary (CHO) cells are a key platform host for the production of therapeutic proteins. Worldwide sales totaled more than $65 billion from the 44 CHO-produced therapeutics in 2012 (Kremkow and Lee 2013; La Merie 2013), accounting for more than 50% of all biopharmaceuticals sales (La Merie 2013; Jayapal et al. 2007). Desirable characteristics of CHO cells that make them advantageous for manufacturing therapeutic proteins include prior Food and Drug Administration

approval, human-like protein modifications (glycosylation and protein-folding), and adaptability to various culture conditions (Wurm 2004; Wuest et al. 2012).

While significant productivity improvements have been made, CHO cell productivities are still significantly lower than bacterial or yeast-derived biopharmaceutical productivities (Jayapal et al. 2007). Transcriptomic, proteomic, and cell engineering approaches (especially gene knock-down studies) were limited until CHO genome sequences became publicly available. Owing to the advent and development of next-generation sequencing technology, CHO host cell lines (CHO-K1 ATCC, CHO-K1 ECACC, and CHO-S), CHO therapeutic production cell lines, and Chinese hamster (CH) cells have been sequenced, and draft CHO-K1 and CH genomes have been assembled and annotated (Xu et al. 2011; Lewis et al. 2013; Brinkrolf et al. 2013). As more host cell and therapeutic production cell lines will likely be sequenced in the near future, there is a need for a central repository for CHO-specific genome information with a comparative gene search function. In addition, providing useful tools for CHO-specific scientific applications, including sequence alignment, genome viewers, and comparative genomic analysis will be beneficial to the CHO community.

The community-wide CHOgenome.org website was launched in 2011 (referred to as version 1.0) and hosted the CHO-K1 mitochondrial genome (Partridge et al. 2007), the CHO-K1 GenBank genome sequence (Xu et al. 2011), and an associated BLAST (Basic Local Alignment Search Tool) service (Hammond et al. 2012). Since then, CHOgenome.org has been improved with the addition of the RefSeq CHO-K1 and CH genome databases, gel-based and shotgun-based proteome databases, and new website features and functions. In this chapter, I describe characteristics of the

different databases, the updated website (version 2.0) features I implemented, and future directions for CHOgenome.org that I believe will aid the CHO community in developing CHO cell-based studies.

3.4 Results and Discussion

3.4.1 Updated Genome and Annotation Databases

3.4.1.1 Annotations Databases

During genome annotation, the annotation pipeline(s) predict genes in silico from a genome assembly despite the predicted results not necessarily being validated by experimental evidence such as RNA-seq data (Yandell and Ence 2012). Therefore, annotated gene information may not be exact and may contain mis-annotations. Moreover, a genome requires ongoing annotation updates when more accurate annotation prediction algorithms, new sequence data (whole genome sequencing), or experimental data (RNA-seq) become available (Brent 2008). GenBank is intended to serve as a primary sequence data archive (National Center for Biotechnology Information 2013); GenBank annotations are user-submitted, can only be curated by the submitter and are seldom updated in practice. Alternatively, the Reference Sequence (RefSeq) database is a collection of curated, non-redundant genome (DNA), transcript (RNA), and protein (amino acid) sequences produced by the National Center for Biotechnology Information (NCBI) (National Center for Biotechnology Information 2013). A RefSeq database is created when nucleotide or amino acid sequences archived in GenBank are independently annotated by the NCBI annotation pipelines. In contrast with GenBank annotations, RefSeq annotations are regularly

updated. CHOgenome.org (version 1.0) contained the CHO-K1 GenBank annotation (Hammond et al. 2012) and CHOgenome.org (version 2.0) will use RefSeq annotations as the standard genome databases when available.

3.4.1.2 *Cricetulus griseus* Genomes

In contrast with other model organisms and primary cells, CHO cells exhibit aneuploidy with a significant number of chromosomal rearrangements caused by genetic instability (Cao et al. 2012; Derouazi et al. 2006). As a result, one cell line's genetic information may be different from other cell lines', resulting in cell line specific genome sequences. Furthermore, changes within the CHO genome occur spontaneously, causing CHO cells to exhibit genomic diversity within a cell line (Lewis et al. 2013; Wurm and Hacker 2011). Given that the CHO-K1 genome (or any CHO host cell line genome) may not serve as the ideal reference genome for all CHO cell lines, recent efforts have been directed towards establishing the CH as the CHO reference genome, as the CH is considered a model organism with stable and intact genome information. Lewis et al. (Lewis et al. 2013) and Brinkrolf et al. (Brinkrolf et al. 2013) separately sequenced and assembled CH genomes. While the Brinkrolf genome draft has chromosome-assigned scaffold information, the Lewis genome draft has better assembly metrics, including a higher sequence coverage, longer scaffold N50 (the length of the smallest scaffold where all scaffolds equal to or longer than that length cover 50% of the genome), and a lower gap ratio. Therefore, the Lewis CH genome draft, as well as the CHO-K1 genome, were chosen by NCBI for RefSeq annotation using the 2014 RefSeq annotation pipeline.

3.4.1.3 RefSeq Annotation

As of September 2016, there are three CH and CHO RefSeq nuclear genome databases on CHOgenome.org (version 2.0): 2012 CHO-K1 RefSeq, 2014 CH RefSeq, and 2014 CHO-K1 RefSeq. Each of the RefSeq databases are unique whole genome annotations. Upon the release of additional databases on CHOgenome.org, each database will remain unchanged, providing users the knowledge of database constancy. Characteristics of the 2012 CHO-K1, 2014 CH, and 2014 CHO-K1 RefSeq genome annotation databases currently hosted on CHOgenome.org are summarized in Figure 3.1 and Table 3.1. There is a large difference in the number of sequences between the 2012 and the 2014 CHO-K1 annotations due to algorithm pipeline improvements. The difference between the 2014 CHO-K1 and 2014 CH annotation statistics are likely due to chromosomal rearrangements and differences in the genome assemblies. Note that the number of sequences in Table 3.1 do not necessarily indicate that shared gene IDs contain identical gene sequences between the databases.

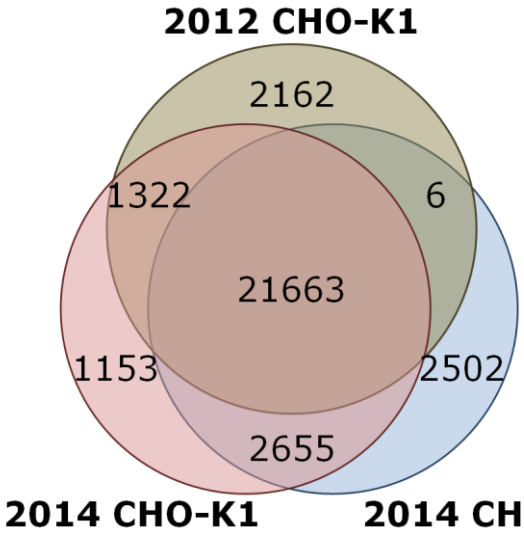


Figure 3.1: CHO-K1 and CH gene annotation comparisons. All genes in the 2012 CHO-K1, 2014 CHO-K1, and 2014 CH RefSeq databases are classified according to the annotation(s) that contain the specific gene IDs. The 2,162 genes found only in the 2012 CHO-K1 database have been discontinued and the associated information is no longer available at NCBI.

Table 3.1: CHO-K1 and CH RefSeq genome database characteristics

Sequence Type	2014 CH	2014 CHO-K1	2012 CHO-K1	2012 vs. 2014 CHO-K1 Differences[a)]	2014 CH vs. CHO-K1 Differences[b)]
Gene	27,545	27,843	25,169	2,674	298
mRNA	29,985	31,545	21,612	9,933	1,560
Exon	336,798	356,321	231,866	124,455	19,523
CDS	302,353	320,307	209,486	110,821	17,954
tRNA	468	501	501	0	33
rRNA	3	3	6	3	0
ncRNA	2,880	2,881	4	2,877	1

a) Differences between the 2012 CHO-K1 RefSeq and 2014 CHO-K1 RefSeq genome annotations
b) Differences between the 2014 CHO-K1 RefSeq and 2014 CH RefSeq genome annotations

3.4.2 Updated Website Features

A dedicated server now hosts the CHOgenome.org databases, increasing reliability and minimizing downtime caused by network dependencies and transition lag. A set of custom-built PHP programs is used to parse the raw data and populate the databases, while the entity-relationship schema was implemented in MySQL. Communication to the databases, query result sorting, and the user interface are also implemented in PHP, as server-side executable functions. I updated the

CHOgenome.org (version 2.0) website's organization and functionality to accommodate additional genome databases, proteome databases, other new features, and to become more user-friendly.

3.4.2.1 Gene Search

3.4.2.1.1 RefSeq Databases

All RefSeq *Cricetulus griseus* genomes can be searched simultaneously in CHOgenome.org (version 2.0), a feature I designed that was not previously available on CHOgenome.org (version 1.0). A comprehensive RefSeq genome search function and an advanced search page with selection options are both available because of the new streamlined database structure. This modified database structure enables users to select which RefSeq genome(s) to search using the official gene symbols, names assigned to the gene products during the NCBI genome annotation, or by the gene's NCBI ID. Temporarily assigned gene symbols are searchable if the gene was not given an official gene symbol during the automated annotation process. These temporary symbols consist of 'LOC' followed by the 9-digit NCBI gene ID (*i.e.* LOC100######).

Genome database search results are listed in a tabular format where each gene's parent genome assembly, feature type, symbol, NCBI ID, and name/product description are displayed. If multiple genomes are searched, all search results are organized by genome. Selection of the gene's NCBI ID provides additional details, including details new to CHOgenome.org (version 2.0), such as protein homologs, transcript homologs, and a related entries table, as shown in Figure 3.2. The human,

mouse, and rat protein homologs have been added to the detailed results webpage and each listed NCBI ID links to the specific gene's NCBI protein webpage. This information was generated by downloading the list of protein homologs of both CHO-K1 and CH from NCBI's HomoloGene database (build 68) and filtering the list to solely contain the human, mouse, and rat entries for each protein. The homolog information will be updated as additional CHO and CH genome annotations are verified. The transcript homologs of the 2014 CHO-K1 and 2014 CH RefSeq databases were assembled by aligning the transcripts with the reciprocal best hits (Wall et al. 2003). All matched transcripts are available in a spreadsheet and upon download, the CHO-K1 or CH homolog can be easily identified using the NCBI ID, gene symbol, transcript ID, or description. The related entries table is another new feature I designed that displays details for every product associated with a gene, including each transcript and protein. Each row represents one feature associated with the gene, where the highlighted row distinguishes the feature whose content is displayed on the current webpage. The columns identify the type of gene product, NCBI transcript/protein ID, amino acid or nucleotide sequence, NCBI graphics, and transcript ID (if multiple transcripts exist for the gene product).

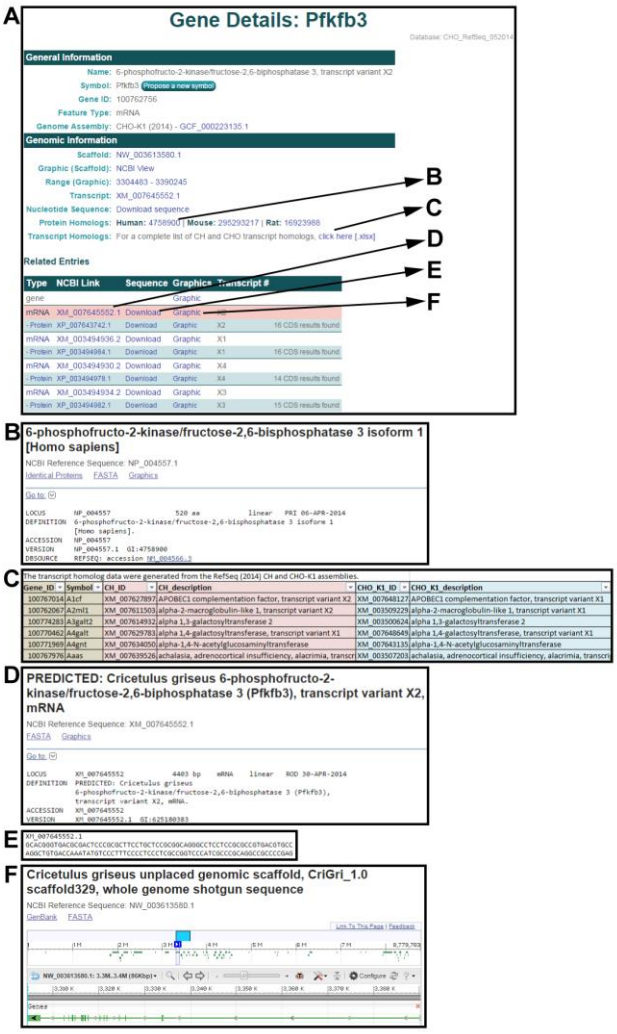


Figure 3.2: CHOgenome.org (Version 2.0) RefSeq gene search result details. (A) Genome database search result details identify each gene's parent assembly, feature type, symbol, NCBI ID, and name/product description, which are displayed in addition to the scaffold on which the gene is located, the coordinates of the gene's coding region, the NCBI transcript ID, a link to the NCBI graphics page, a link to the FASTA nucleotide or amino acid sequences, links to the NCBI protein homologs in human, mouse, and rat, *Cricetulus griseus* transcript homologs, and a related entries table. The new features I implemented are the (B) protein homologs, (C) transcript homologs, and related entries table, consisting of the gene product type, (D) NCBI reference, (E) nucleotide/amino acid sequence, (F) NCBI graphic, and transcript # (if appropriate). The arrows identify links to new features and the letters correspond to the webpage or information each link provides.

3.4.2.1.2 Original CHO-K1 Mitochondrial and GenBank Genome Database

Separate search pages for both the CHO-K1 mitochondrial and GenBank CHO-K1 genomes are also available on CHOgenome.org (version 2.0), similar to the CHOgenome.org (version 1.0) search pages. The CHO-K1 mitochondrial genome information is included in the RefSeq CHO-K1 genome assemblies, but can be independently searched in a separate CHOgenome.org database. The 2011 CHO-K1 GenBank genome can be searched by accession number, gene name or symbol, and Gene Ontology (GO) term. While the results are similar to the RefSeq search results, the 2011 CHO-K1 GenBank genome database information differs from the RefSeq CHO-K1 genome assemblies because the GenBank database contains GO terms, fewer gene symbols, and fewer NCBI IDs.

3.4.2.2 BLAST

The CHO BLAST performs alignments of nucleotide or amino acid sequences against the CHO and CH genome sequence databases. Currently, the CHO BLAST server for CHOgenome.org (version 2.0) has 13 nucleotide and amino acid databases, an increase of seven from version 1.0. The organization of these databases has been greatly improved upon, as the nucleotide databases are divided into genome (scaffold) and transcript (RNA) databases, while the amino acid databases consist of protein databases. Single or multiple query sequences can be entered in FASTA format into the search box or uploaded as a FASTA file and searched against one or multiple databases concurrently. BLAST search results are summarized in a table according to the query sequence name, subject sequence name, bit score, identity length, identity

percentage, and E-value. The results can be filtered by score, similarity cutoff percentage, or BLAST bit score (Fassler and Cooper 2008). Additional features include viewing the pair-wise alignment, downloading the BLAST sequences, and viewing the RefSeq/GenBank entry for each result. Below the required parameters, additional settings can be changed to perform a more advanced BLAST search.

3.4.2.3 Genome Viewer

The CH and CHO-K1 genomes can be viewed through the JBrowse genome viewer tool (Skinner et al 2009), a novel tool for CHOgenome.org (version 2.0). JBrowse is an open-source, embeddable genome browser built with JavaScript and HTML5, and contains easy-to-use Perl scripts to format data. The data files are then read over HTTP, leaving JBrowse very light on the back-end and able to handle large (multi-gigabase) genomes. This arrangement allows users fast and smooth viewing of the genome and its annotations.

The genome viewer currently provides visualization of the 2014 annotated CHO-K1 genome and the chromosome-sorted CH genome (Brinkrolf et al. 2013). Tracks for genes, mRNA, exons, GC-content, and other annotation data are available for the CHO-K1 genome. The CH genome viewer is in its beta-testing stage and does not currently have annotations uploaded, but a user can add tracks containing their own annotation data. Once tracks have been added, JBrowse allows for the merging of multiple tracks into one containing the tracks' intersection or union. Several accepted track data formats are GFF3 (Reese et al. 2010), VCF (Danecek et al. 2011), BAM (Li et al. 2009), and BigWig (Kent et al. 2010).

3.4.2.4 Other Resources

The proteome browser is a novel tool I adapted to provide access to public proteomic data, specifically two-dimensional polyacrylamide gel electrophoresis (2D PAGE) gel and shotgun proteomic data generated from CHO cell lines (Hayduk et al. 2004). The 2D PAGE gel feature lists all submitted 2D gels in the database with a small image of the gel and the publication information (if applicable). Once selected, a table of the identified proteins with links to the respective NCBI protein and CHOgenome.org gene pages, an enlarged gel image, and a link to the reference are provided. The shotgun proteomics database is searchable by protein name and consists of a description of each dataset and the publication of origin. The search results are comprised of the protein name and accession identification, as well as other details, such as the identified peptide sequences, number of identified sequences, coverage, false discovery rate, cluster ID, group ID, SwissProt Annotation Homology, GO annotation, and sequence ID.

The file archive is a novel resource to CHOgenome.org (version 2.0) I helped implement to host all CHOgenome.org files, including all current and outdated CHO and CH files. The CHOgenome.org file archive serves as a repository for the entire CHO community, and the CHOgenome.org file archive files are never removed. The files are organized by the organism or cell line name and sequencing project, and consist of annotation and sequence files.

CHOgenome.org genome databases are generated after new or updated *Cricetulus griseus* genomic information is published or released. Upon the posting of an updated CHOgenome.org database that replaces another database, the previous database will be reclassified as a legacy database. Only one legacy database for each

Cricetulus griseus cell type or cell line will be maintained; upon the classification of a new legacy database for one cell type or cell line, the older legacy database will be removed from CHOgenome.org. Upon removal of a searchable genome database from CHOgenome.org, notice will be given to the CHOgenome.org community and the older legacy database will be removed. However, the older legacy database's annotation and sequence source files will remain in the file archive. This process promotes use of the most recent information, while maintaining outdated information in a manner that benefits the entire CHO community. Similarly, when significant upgrades are made to the website itself, the previous version of the website will be temporarily maintained, permitting users a transition period to the updated website. The file archiving process is comparable to the process used by many other genome websites (Suresh et al. 2014; Blake et al. 2014; McQuilton et al. 2012; Laulederkind et al. 2013), yet the maintenance of legacy databases is less common (Laulederkind et al. 2013).

3.4.3 Impacts of CHOgenome.org on the CHO Community

One major benefit of using CHOgenome.org is that a user can obtain CHO and CH specific sequence information. As RefSeq will ultimately contain only one sequence per gene for each organism, sequence variants resulting from different CHO cell lines will not be represented at NCBI. For example, the sequence comparison of the tumor protein p53 (*Tp53*) gene between the CHO-K1 and CH genomes identifies four amino acid differences in the coding region. However, there is only one *Cricetulus griseus* RefSeq mRNA sequence for the *Tp53* gene (Accession no. NM_001243976.1) at NCBI, the sequence of which corresponds to the CH *Tp53* gene.

This lack of sequence information may limit the design of site-specific engineering approaches, such as CRISPR/Cas9 and transcription activator-like effector nucleases (TALEN), where exact sequence information is critical. However, CHOgenome.org contains all four *Tp53* sequences because CHOgenome.org has sequence information (genes, transcripts, and protein sequences) for each sequenced cell line and will host the sequences for additional host cell lines. The goal is to enable users to search and obtain all published sequences of interest and determine which is most accurate for their application or cell line.

CHOgenome.org has enabled access to CHO-specific genome information to the community for broad use, likely increasing the number of published experimental applications. For example, the CHO BLAST tool has been used to determine the similarity of mouse and human miRNA sequences to CHO sequences for use in CHO differential miRNA expression studies (Maccani et al. 2014). The goal of this study was to identify miRNAs that are involved in heterologous protein synthesis and secretion by investigating the miRNA expression patterns of high, low, and non-producing recombinant CHO cell lines. The results indicate that CHO cell heterologous protein expression effects are strongly product- and/or clone-specific. The CHOgenome.org gene search function has been used to identify nucleotide mismatches between experimental DNA assemblies and GenBank records (Orlova et al. 2012). The goal of this study was to develop an approach that efficiently designs and constructs a plasmid clone for a DNA fragment larger than 5 kilobase pairs (kbp). Two non-coding fragments of the CHO translation elongation factor 1 alpha gene were used to validate the computer tool-based modular DNA assembly. The CHO genome has also enabled the creation of new bioinformatics tools specifically

targeting the CHO community, including a web-based target-finding tool that identifies potential target sequences applicable for the CRISPR/Cas9 system (Ronda et al. 2014), which has been cited by sixteen non-review scientific publications in less than two years. These few examples demonstrate how CHOgenome.org's features are being used, in addition to spurring the creation of new tools for use within the CHO community.

The estimated number of CHOgenome.org users since 2011 is at least 20,000, originating from more than 1,700 different cities in 84 countries on six different continents, as shown in Figure 3.3. The city distribution is uneven, as the majority of users are located in Europe, the United States, and Asia. The five countries with the largest number of sessions are the United States, Germany, the United Kingdom, Japan, and China, respectively, accounting for nearly 65% of all measured traffic.

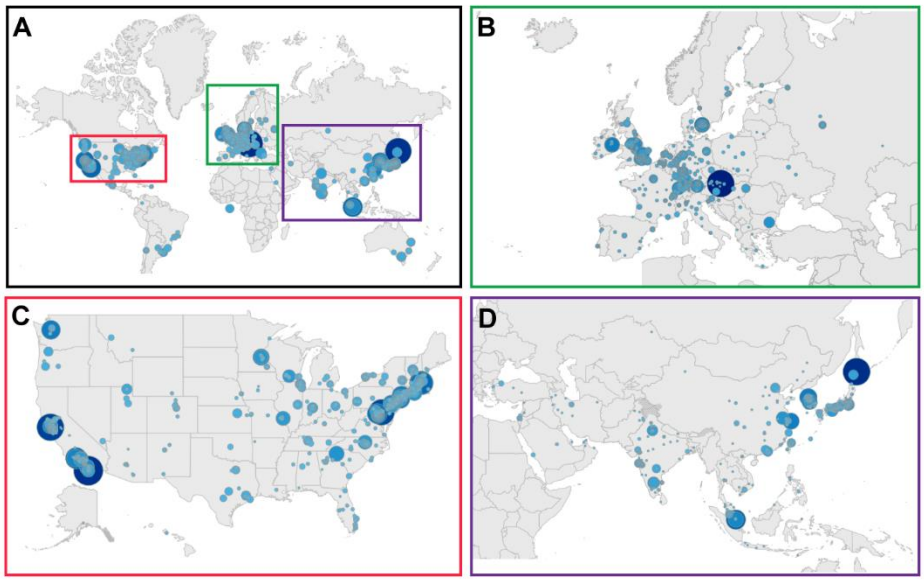


Figure 3.3: CHOgenome.org world usage map. Each circle identifies a city with at least three different users of CHOgenome.org from 2011 through 2014, and the circles are proportional (in size and shading) to the number of users. The colored rectangles in (A) the world usage map correspond to the zoomed images of (B) Europe, (C) the United States, and (D) Asia. These maps exclude Newark, DE, USA, where the CHOgenome.org server is located.

3.5 Concluding Remarks and Future Directions

The CHO genome is arguably heterogeneous and is certainly ever-changing in culture, thus any given version of the CHO genome does not serve as a suitable reference genome for the CHO community. Therefore, efforts to improve the quality of the CH genome and to make the CH genome the *Cricetulus griseus* reference genome have begun. Pacific Biosciences (PacBio) sequencing (Eid et al. 2009) of CH samples is currently underway at the University of Delaware and the Johns Hopkins University sequencing centers. The sequence reads produced by PacBio (>6 kbp)

(Chaisson et al. 2015) are much longer than short read sequencing technology reads (150-600 base pairs) (Kremkow and Lee 2015). This increased read length will be useful as a large number of contigs and scaffolds still need to be ordered and linked together in the current CH draft genomes. PacBio sequencing reads will facilitate this effort, while also filling gaps within the scaffolds. A more complete CH reference genome will enable better genome annotation and comparisons between CHO cell lines.

It is anticipated that more CHO cell line genomes will be sequenced, annotated, and publicly released in the future and CHOgenome.org plans to host the data, enabling researchers quick access to all CHO genomic information. Additionally, hosting comparative genomic information not only at the sequence level (single nucleotide variants and short insertions/deletions), but also at the chromosome level (structural variations such as large insertions/deletions, inversions, translocations, and copy number variations) will provide cell line-specific genomic variations for the CHO community. An efficient visualization tool should be developed to display comparative analysis results for multiple genomes. In addition, small RNAs (such as small interfering RNA, miRNA, or single guide RNA of the CRISPR/Cas9 system) are increasingly recognized as engineering candidates with advantages including sequence specific activity and low metabolic burden (Ronda et al. 2014; Zamore and Haley 2005; Baik and Lee 2014; Klanert et al. 2014). A cell line specific small RNA design tool would support an increase in small RNAs research.

To improve CHO bioprocessing, the CHO cell physiology must be better understood. While the CHO-K1 & CH genomes reveal an abundance of information, other -omics analyses, such as transcriptomics, proteomics, and metabolomics, are

often more relevant, as these studies provide different portraits of the underlying systems biology. Mammalian cell genomics, transcriptomics, proteomics, and metabolomics have been studied in great detail over the past few years (Wuest et al. 2012; Valente et al. 2014; Datta et al. 2013; Vishwanathan et al. 2014; Heffner et al. 2014; Dickson 2014; Dietmair et al. 2012; Kim et al. 2012; Kildegaard et al. 2013). However, the rare linearity between these different –omics datasets indicates that each data type does not solely represent the complete biological profile for a given CHO cell physiology.

The simultaneous use of multiple –omics technologies for one experiment can provide additional data and possibly a greater understanding of CHO cell physiology. However, designing experiments that apply multiple -omics tools and interpreting the variably formatted datasets accurately can be challenging. Studies using multiple –omics technologies are beginning to be published and a few reviews explore the current state of using multiple –omics technologies for CHO research (Kremkow and Lee 2013; Farrell et al. 2014). While CHOgenome.org will continue to host all CHO genomic data, one long-term CHO community objective is to integrate available genomic, transcriptomic, proteomic, metabolomic, and multiple –omics datasets. The creation of the CHO proteome database represents initial steps toward incorporating nongenome CHO data. A public, extensive CHO –omics data archive will hopefully lead to the generation of genome-scale models, a comprehensive understanding of CHO cellular behavior, and possible engineering targets and bioprocessing improvements.

3.6 Website and Partners Information

The complete URL for the CHO genome website is http://www.chogenome.org/. Questions and comments can be sent to chogenome@dbi.udel.edu. CHOgenome.org is first and foremost a community resource, and the utility, longevity, and success of CHOgenome.org depends upon community support. Partners of CHOgenome.org include government agencies, corporations, and universities actively supporting the CHO genome initiative. Representatives from each of the partners are invited to participate in the steering committee to discuss updates to the data and the website and each partner may choose to have a logo on the CHOgenome.org website.

3.7 Acknowledgements

I am grateful for financial support from the National Science Foundation (1412365, 1144726, and 1247394), the National Institute of Standards and Technologies (60NANB11D185), as well as from Amgen, Biogen Idec, Genentech, GT Life Sciences, Life Technologies, University of Queensland, Austrian Centre of Industrial Biotechnology, BOKU University, Danish Technical University, Dublin City University, University of California San Diego, Johns Hopkins University sequencing center, and the University of Delaware sequencing center. I thank Shawn Polson, Blake Meyers, Mike Betenbaugh, Nicole Borth, Cathy Wu, Eric Garrison, and Karol Miaskewicz for important discussions.

REFERENCES

Baik JY, Lee KH. (2014) miRNA expression in CHO:Nature knows best. Biotechnol J. 9:459-460.

Baycin-Hizal D, Tabb DL, Chaerkady R, Chen L, Lewis NL, Nagarajan H, Sarkaria V, Kumar A, Wolozny D, Colao J, Jacobson E, Tian Y, O'Meally RN, Krag SS, Cole RN, Palsson BØ, Zhang H, Betenbaugh M. (2012) Proteomic analysis of Chinese hamster ovary cells. J Proteome Res. 11:5265-5276.

Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE, The Mouse Genome Database Group. (2014) The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res. 42:D810-817.

Brent, MR. (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet. 9:62-73.

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

Cao Y, Kimura S, Itoi T, Honda K, Ohtake H, Omasa T. (2012) Construction of BAC-based physical map and analysis of chromosome rearrangement in Chinese hamster ovary cell lines. Biotechnol Bioeng. 109:1357-1367.

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J, Eichler EE. (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature. 517:608-611.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. (2011) The variant call format and VCF tools. Bioinformatics. 27:2156-2158.

Datta P, Linhardt RJ, Sharfstein ST. (2013). An ‘omics approach towards CHO cell engineering. Biotechnol Bioeng. 110:1255-1271.

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

Dickson AJ. (2014) Enhancement of production of protein biopharmaceuticals by mammalian cell cultures: the metabolomics perspective. Curr Opin Biotechnol. 30:73-79.

Dietmair S, Nielsen LK, Timmins NE. (2012) Mammalian cells as biopharmaceutical production hosts in the age of omics. Biotechnol J. 7:75-89.

Eid J, Fehr A, Gray J, Luong K, Lyle F, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, deWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, et al. (2009) Real-time DNA sequencing from single polymerase molecules. Science. 323:133-138.

Farrell A, McLoughlin N, Milne JJ, Marison IW, Bones J. (2014) Application of multi-omics techniques for bioprocess design and optimization in Chinese hamster ovary cells. J Proteome Res. 13:3144-3159.

Fassler J, Cooper P. (2008) National Center for Biotechnology Information (US) (Ed.), BLAST® Help, National Center for Biotechnology Information (US), Bethesda. BLAST Glossary.

Hammond S, Kaplarevic M, Borth N, Betenbaugh MJ, Lee KH. (2012) Chinese hamster genome database: an online resource for the CHO community at www.CHOgenome.org. Biotechnol Bioeng. 109:1353-1356.

Hayduk EJ, Choe LH, Lee KH. (2004) A two-dimensional electrophoresis map of Chinese hamster ovary cell proteins based on fluorescence staining. Electrophoresis. 25:2545-2556.

Heffner KM, Hizal DB, Kumar A, Shiloach J, Zhu J, Bowen MA, Betenbaugh MJ. (2014) Exploiting the proteomics revolution in biotechnology: from disease and antibody targets to optimizing bioprocess development. Curr Opin Biotechnol. 30:80-86.

Jayapal KP, Wlaschin KF, Hu WS, Yap MGS. (2007) Recombinant protein therapeutics from CHO cells - 20 years and counting. Chem Eng Prog. 103:40-47.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 26:2204-2207.

Kildegaard HF, Baycin-Hizal D, Lewis NL, Betenbaugh MJ. (2013) The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology. Curr Opin Biotechnol. 24:1102-1107.

Kim JY, Kim Y, Lee GM. (2012) CHO cells in biotechnology for production of recombinant proteins: current state and further potential. Appl Microbiol Biotechnol. 93:917-930.

Klanert G, Jadhav V, Chanoumidou K, Grillari J, Borth N, Hackl M. (2014) Endogenous microRNA clusters outperform chimeric sequence clusters in Chinese hamster ovary cells. Biotechnol J. 9:538-544.

Kremkow B, Lee KH. (2013) Next-generation sequencing technologies and their potential impact on CHO cell-based biomanufacturing. Pharm Bioprocess. 1:455-465.

Kremkow BG, Lee KH. (2015) Sequencing technologies for animal cell culture research. Biotechnol Lett. 37:55-65.

La Merie Business Intelligence. (2013) Blockbuster biologics 2012. R&D Pipeline News. 7:2-28.

Laulederkind SJ, Hayman GT, Wang SJ, Smith JR, Lowry TF, Nigam R, Petri V, de Pons J, Dwinell MR, Shimoyama M, Munzenmaier DH, Worthey EA, Jacob HJ. (2013) The Rat Genome Database 2013 – data, tools, and users. Brief Bioinform. 14:520-526.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078-2079.

Maccani A, Hackl M, Leitner C, Steinfellner W, Graf AB, Tatto NE, Karbiener M, Scheideler M, Grillari J, Mattanovich D, Kunert R, Borth N, Grabherr R, Ernst W. (2014) Identification of microRNAs specific for high producer CHO cell lines using steady-state cultivation. Appl Microbiol Biotechnol. 98:7535-7548.

McQuilton P, St Pierre SE, Thurmond J, FlyBase Consortium. (2012) FlyBase 101--the basics of navigating FlyBase. Nucleic Acids Res. 40:D706-714.

National Center for Biotechnology Information (US). (2013) The NCBI Handbook. National Center for Biotechnology Information (US), Bethesda.

Orlova NA, Orlov AV, Vorobiev II. (2012) A modular assembly cloning technique (aided by the BIOF software tool) for seamless and error-free assembly of long DNA fragments. BMC Res Notes. 5:303.

Partridge MA, Davidson MM, Hei TK. (2007) The complete nucleotide sequence of Chinese hamster (Cricetulus griseus) mitochondrial DNA. DNA Seq. 18:341-346.

Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, Marth GT, Stein L, Flicek P, Yandell M, Eilbeck K. (2010) A standard variation file format for human genome sequences. Genome Biol. 11:R88.

Ronda C, Pedersen LE, Hansen HG, Kallehauge TB, Betenbaugh MJ, Nielsen AT, Kildegaard HF. (2014) Accelerating genome editing in CHO cells using CRISPR Cas9 and CRISPy, a web-based target finding tool. Biotechnol Bioeng. 111:1604-1616.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. (2009) JBrowse: a next-generation genome browser. Genome Res. 19:1630-1638.

Suresh BV, Roy R, Sahu K, Misra G, Chattopadhyay D. (2014) Tomato genomic resources database: an integrated repository of useful tomato genomic information for basic and applied research. PLoS One. 9:e86387.

Valente KN, Schaefer AK, Kempton HR, Lenhoff AM, Lee KH. (2014) Recovery of Chinese hamster ovary host cell proteins for proteomic analysis. Biotechnol J. 9:87-99.

Vishwanathan N, Le H, Le T, Hu WS. (2014) Advancing biopharmaceutical process science through transcriptome analysis. Curr Opin Biotechnol. 30:113-119.

Wall DP, Fraser HB, Hirsh AE. (2003) Detecting putative orthologs. Bioinformatics. 19:1710-1711.

Wuest DM, Harcum SW, Lee KH. (2012) Genomics in mammalian cell culture bioprocessing. Biotechnol Adv. 30:629-638.

Wurm FM, Hacker D. (2011) First CHO genome. Nat Biotechnol. 29:718-720.

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

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

Yandell M, Ence D. (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 13:329-342.

Zamore PD, Haley B. (2005) Ribo-gnome: the big world of small RNAs. Science. 309:1519-1524.

Chapter 4

GLYCO-MAPPER DEVELOPMENT AND GLYCOFORM PREDICTION CAPABILITY VALIDATION

4.1 Preface

This chapter is adapted from Kremkow and Lee (2016-Submitted) with permission in accordance with the guidelines in the publishing contract. This chapter presents the novel modeling technique I created titled Discretized Reaction Network Modeling with Fuzzy Parameters (DReaM-zyP) and the application of DReaM-zyP to create the CHO-specific glycoform prediction tool titled Glyco-Mapper. Glyco-Mapper predictions were validated using a variety of published cell engineering CHO glycosylation studies representing the many different glycoform engineering strategies. A novel cell engineering strategy was predicted and experimentally validated. The impact this modeling technique and this glycoform prediction tool will have on CHO cell-based biomanufacturing applications are analyzed. MALDI mass spectrometry operation was performed by Leila Choe.

4.2 Abstract

Glyco-Mapper is a novel systems biology product quality prediction tool created using a new framework termed: Discretized Reaction Network Modeling using Fuzzy Parameters (DReaM-zyP). Within Glyco-Mapper, users control the nutrient feed composition and the reaction fluxes of glycosylation genes to match a reference glycoform, enabling cell-line specific glycoform predictions as a result of specific nutrient feeding and cell engineering strategies. Glyco-Mapper accurately predicts all published genetically altered glycoforms between 1999 and 2014 with an accuracy, sensitivity, and specificity of 96%, 85%, and 97%, respectively. The modeled glycoforms span a large range of glycoform engineering strategies, including the altered expression of glycosylation, nucleotide sugar transport, and metabolism genes, as well as an altered nutrient feeding strategy. A glycoprotein-producing CHO cell line reference glycoform was modeled and a novel Glyco-Mapper prediction was experimentally confirmed with a respective accuracy and specificity of 95% and 98%. Glyco-Mapper, a product quality prediction tool provides a streamlined way to design host cell line genomes to achieve specific product quality attributes.

4.3 Introduction

Chinese hamster ovary (CHO) cells accounted for the production of $85 billion of the $154 billion in global biotherapeutic sales in 2015 (LaMerie, 2016) because CHO cells are capable of producing large amounts of protein, sustaining high viability, resisting viral infection, and mimicking human product quality attributes (Jayapal et al., 2007; Xu et al., 2011). The protein product quality largely influences

the efficacy, half-life, and immunogenicity of the therapeutic protein (Walsh and Jefferis, 2016), all of which greatly affect the patient's clinical response. The 2013 biopharmaceutical pipeline contained 431 recombinant proteins and monoclonal antibodies (mAbs) in various phases of development (ABRC, 2013). Coupling this large number of pipeline products with the desire to meet quality guidelines in accordance with FDA guidance (US DHHS FDA, 2015), Understanding and predicting the therapeutic product quality in accordance with targeted modifications will be greatly beneficial to biopharmaceutical production.

Several kinetic and data-dependent models have previously been published with the goal of establishing frameworks to quantify and model glycosylation. The Umaña model (Umaña and Bailey, 1997) mathematically depicted the glycosyltransferase activity of 8 enzymes, 33 species, and 33 reactions. Krambeck (Krambeck and Betenbaugh, 2005) expanded upon the Umaña model by incorporating more variables and models 11 enzymes, 7,565 species, and 22,871 reactions, increasing the model's complexity. Krambeck (Krambeck et al., 2009) further broadened the model to incorporate 19 enzymes, more than 10,000 species, and generates and optimizes a synthetic mass spectrum to determine the likely enzyme concentrations. Liu (Liu and Neelamegham, 2014) further developed the glycan mass spectra analysis to construct biochemical reaction networks and calculate the associated fluxes for both N- and O-glycosylation associated pathways, in addition to determining enzyme activities. Spahn (Spahn et al., 2016) employed Markov chain modeling to mathematically calculate parameters to reproduce various glycoform distributions and does not require user-provided kinetic information. Despite the power of current analytical methods, kinetic model parameter estimation and

validation is difficult to achieve for all relevant enzymes using current experimental techniques. Moreover, measurement of glycan stereoisomers and confirmation of a glycan's specific production pathway is not yet possible on a routine basis.

In contrast to detailed kinetic and data-driven models, genome-scale reconstructions can model biological processes and have been successfully applied towards mammalian systems in many contexts (Duarte et al., 2007; Selvarasu et al., 2010; Shlomi et al., 2008) and are now being used to study CHO (Chen et al., 2012; Chowdhury et al., 2015; Selvarasu et al., 2012). The current availability of the CHO genome (Xu et al., 2011) offers an opportunity to investigate CHO-specific product quality using a genome scale reconstruction-based model. This type of model has reduced computational requirements compared to detailed kinetic models, but does not fully capture the variety of products generated by a non-template driven process. Here we report for the first time a modeling framework rooted in genome reconstruction but using discretized parameters and reaction stoichiometry termed Discretized Reaction Network Modeling using Fuzzy Parameters (DReaM-zyP) to predict the likely products of a non-template driven process, specifically the glycosylation patterns of therapeutic proteins. A specific glycosylation DReaM-zyP-based tool (Glyco-Mapper) includes all CHO N-glycosylation genes, as well as nucleotide sugar synthesis, transporter, and glycosylation-relevant metabolism genes. Glyco-Mapper models and predicts the published cell-engineered glycoforms from 1999 to 2014 (Goh et al., 2014; Imai-Nishiya et al., Kanda et al., 2007; Malphettes et al., 2010; Maszczak-Seneczko et al., 2013; Naso et al., 2010; Onitsuka et al., 2012; Sealover et al., 2013; 2007; Tsukahara et al., 2006; Weikert et al., 1999) with an accuracy of 96% and is

currently implemented in Microsoft Excel, which does not require the user to have extensive knowledge of modeling software or programming.

4.4 Materials and Methods

4.4.1 Model Glycosylation and Metabolism Gene Sources

Glyco-Mapper contains 59 N-linked glycosylation genes (Table E.1) and 92 metabolism-related genes (Table E.2), encompassing the central carbon metabolism (CCM), sugar nucleotide synthesis, and sugar nucleotide transporter pathways (Figure E.1). The genes were manually verified to be present within the CHO and Chinese hamster genomes (Xu et al. 2011; Lewis et al. 2013; Brinkrolf et al. 2013) and the gene sequences were obtained from the 2014 RefSeq CHO-K1 and CH genome annotations (Hammond et al. 2012; Kremkow et al. 2015). Most of the N-linked glycosylation genes were obtained from the CHO-K1 genome sequencing publication (Xu et al. 2011) and additional genes were added to the model from literature (Bosques et al. 2010). The N-linked glycosylation enzyme functions span the entire N-glycosylation reaction network and range from the production of the glycan intermediate in the endoplasmic reticulum to the degradation of the glycan outside of the Golgi. The metabolism-related genes were identified from published CHO CCM models (Ahn and Antoniewicz 2012), as were the genes involved with the sugar nucleotide production and nucleotide sugar transport pathways. Literature as well as the KEGG database were used to define the reactants, products, and enzymatic reaction conditions (Taniguchi et al. 2002; Kanehisa and Goto 2000). Glyco-Mapper models 448 non-stereospecifc glycans representing more than 2,600 distinct,

stereospecifc glycans. Stereoisomers are considered identical for this work because most current analytical glycan methods do not distinguish between stereospecific glycans and there is no reported link between stereoisomers and biotherapeutic characteristics.

4.4.2 Equations, Inputs, and Outputs

A draft reconstruction of the glycosylation process and CCM and nucleotide sugar transport systems was created using the CHO-K1 and Chinese hamster genome annotations, comprehensive reaction databases for candidate metabolic functions, and published experimental data. The glycosylation genes were organized into functional classes, the CHO-specific metabolic map (Figure E.1) was defined, and substrate and cofactor usage, neutral enzymatic reactions, gene and reaction localization, heteromeric enzyme complexes, isozyme functionalities, intracellular transport mechanisms, and supporting metabolic reactions were all verified. The mathematical model was created using fundamental kinetics requiring media components and a kinetic activity level value (k_{ALV}) parameter summarizing each enzyme's activity and concentration, based upon fuzzy logic discrete, quantized variables determining the likely glycoform composition. The Glyco-Mapper k_{ALV} emulate the original gene and enzyme levels (if known), but if unknown, are set to minimize the number of differences between the experimental reference glycoform and the resulting Glyco-Mapper glycoform. Unbalanced reactions, missing reactions, and reaction directionality and limitations were identified and rectified, enabling testing of single-gene (or multi-gene) alteration phenotypes for comparison with experimental data.

Glyco-Mapper inputs include the type of recombinant protein [mAb or non-mAb] and a cellular location of the glycoprotein [secreted or intracellular]. Both parameters restrict the potential glycoform based on respective limitations. The k_{ALV} for each glycosylation and metabolism gene is an input variable that accounts for both the expression level of the gene and the activity of the corresponding enzyme, ranging in value between 0 and 5, to determine the potential glycoform composition. The media sugar component list input enables sugar nucleotide metabolism calculations.

Glyco-Mapper outputs include a count and list of the predicted glycoform glycan composition. Glyco-Mapper generates four glycoform lists dependent upon two different parameters, glycan classification [individual glycans or glycan nucleotide groupings] and kinetic classification [likely secreted glycoform or comprehensive intracellular glycoform composition]. Each glycoform version yields a slightly different view and understanding of the glycosylation reaction network. Lastly, an optional user-selected glycan is predicted to be present or absent in the final glycoform, and if absent, the metabolism or glycosylation genes preventing the glycan's production are identified.

4.4.3 Experimental

I adapted a CHO-DUKX cell line producing secreted alkaline phosphatase (SEAP) (Hayduk and Lee 2005) to serum-free, suspension culture in 125 mL shake flasks (Corning, Oneonta, NY) containing 28 mL SFM4CHO medium (Hyclone Laboratories Inc., Logan, UT). The cells were cultured by routine passaging at 4 day intervals. Cultures were then seeded at $3x10^5$ cells/mL and incubated with orbital agitation at 120 rpm in a 37 °C cell culture incubator with 5% CO_2 and 80% relative

humidity. Cells were counted using a Countess II FL hemocytometer (ThermoFisher, Rockford, IL) with viability determined by Trypan blue (Sigma-Aldrich, St. Louis, MO) exclusion method. The cells were harvested on day 3 or 4 and the supernatant was separated from the residual cells by centrifugation (180 *g*, 6 min) and stored at -20°C until further use.

Supernatant samples were thawed simultaneously and filtered through a 0.22 µm filter (Millipore, Cork, Ireland). A SEAP-activity assay (ThermoFisher, Rockford, IL) was performed on all supernatant samples to quantify the SEAP protein concentration. SEAP was purified using a Reactive Green 19 pseudo-affinity chromatography column, generated according to the published protocol (Ouyang et al. 2007). Briefly, Sepharose^{TM} 6B (GE Healthcare, Uppsala, Sweden) is hydrated and reacted with Reactive Green 19 (Sigma-Aldrich, St. Louis, MO), Na_2CO_3 (Sigma-Aldrich, St. Louis, MO), and 20% NaCl (Fisher, Fair Lawn, NJ), incubated for 48 hours, and thoroughly rinsed with deionized water. The matrix is equilibrated in ethanolamine (Sigma-Aldrich, St. Louis, MO) for 12 hours, rinsed with water, and stored at 4 °C. SEAP is loaded to the column for 8 hours, washed with Tris buffer (Bio-Rad, Hercules, CA), eluted with Na_2HPO_4 buffer (Fisher, Fair Lawn, NJ), and the column is regenerated. The purified SEAP concentration was again measured by the SEAP-activity assay and 200 µg of SEAP per sample was concentrated using 10 kDa centrifugation filters (Waters, Boston, MA) for the permethylation assay. Briefly, SEAP was denatured and digested with trypsin (Promega, Madison, WI) and the glycans were cleaved by N-glycanase (ProZyme, Hayward, CA). The cleaved glycans were purified with Hypersep Hyper Carb SPE cartridges (ThermoFisher, Rockford, IL) using 5% v/v acetonitrile with 0.1%v/v TFA as a wash and 50% acetonitrile with

0.1% v/v TFA to elute the glycans. The elution solution was evaporated under airflow and the glycans were reconstituted and permethylated using methyl iodide in the presence of NaOH and DMSO. The permethylated glycan samples were first cleaned up using liquid-liquid extraction with chloroform and then Sep-Pak PS2 SPE cartridges (Waters, Milford, MA) with elution fractions in 15%, 35%, 50%, and 75% acetonitrile. Eluted fractions were evaporated with a vacuum concentrator, then resuspended in 25 µL of 80% methanol. MALDI TOF/TOF glycan analysis was performed with 10,000 shots at 5000 laser power in positive ion reflector mode with 2,5-dihydroxybenzoic acid matrix using a 4800 MALDI TOF/TOF mass spectrometer (ABSciex, Framingham, MA). The relative glycan percentage was determined as the ratio of the individual glycan peak height to the sum of all glycan peak heights (Figures F.1 and F.2).

GnT-II knockdown was performed using transfection of *GnT-II* (CGAAUACCCUGACUCCUUUdTdT) and negative control #1 siRNA (Sigma-Aldrich, St. Louis, MO) using Lonza transfection Cell Line Nucleofector Kit V (Lonza, Basel, Switzerland). Cells were cultured for 3 days before the supernatant was collected. *GnT-II* knockdown was confirmed by qRT-PCR (Figure F.3) using the TaqMan® RNA-to-Ct 1-Step Kit (Applied Biosystems, Foster City, CA), probe [PrimeTime 5’ 6-FAM/ZEN/3’ IBFQ], and primers [5’-GGGCATTAACGAAGTCCTAGTC-3’; 5’-CAGCTGAATGCTGAATGGAAAG-3’] (IDT, Coralville, IA). qRT-PCR was performed in triplicate on a Cepheid SmartCycler II (Cepheid, Sunnyvale, CA).

4.4.4 Statistical Information

For each predicted glycoform, the accuracy, specificity, and sensitivity statistics are solely representative of the predicted glycoforms, not the reference glycoforms or the combination thereof. The accuracy percentage represents the percentage of true glycan predictions within the glycoform and was calculated as the sum of experimentally-validated, present and absent glycans predicted divided by the total number of glycans within the glycoform. The specificity percentage represents the true negative prediction rate, specifically calculated as the number of predicted and experimentally absent glycans divided by the total number of experimentally absent glycans. The sensitivity percentage represents the true positive prediction rate, specifically calculated as the number of both predicted and experimentally present glycans divided by the total number of experimentally present glycans. The delta accuracy percentage represents the accuracy rate of the glycans that changed their combination of prediction and experimental status between the reference and predicted glycoforms, specifically calculated as the number of correctly predicted glycans that changed status divided by the total number of glycans that changed status. The relative composition deemed statistically significant was greater than 1% for all literature and experimental calculations.

The *GnT-II* siRNA knockdown qRT-PCR data was analyzed as technical triplicates of biological triplicates (Figure F.3). The results were statistically analyzed in JMP and the two sample one-sided t-test was conducted assuming unequal variances with an alpha of 0.05 and resulting in $p<0.0001$. The difference between the *GnT-II* knockdown and negative control samples was 2.51, the t ratio was 7.5, the standard error difference was 0.335, and the degrees of freedom was 15.4.

4.5 Results

Various metabolic and glycosylation gene knockouts, knockdowns, or overexpressions, which will be referred to as cell- or glycoform-engineering in this report, have been described in different CHO cell lines and each of these changes results in altered recombinant protein glycoforms. Between 1999 and 2014, ten publications (Goh et al., 2014; Imai-Nishiya et al., Kanda et al., 2007; Malphettes et al., 2010; Maszczak-Seneczko et al., 2013; Naso et al., 2010; Onitsuka et al., 2012; Sealover et al., 2013; 2007; Tsukahara et al., 2006; Weikert et al., 1999) describe an engineered change in glycosylation-related gene expression with an accompanying characterization of the resulting glycoform changes. These papers collectively altered nine genes affecting eight different sugar nucleotide enzymatic reactions in various combinations among CHO cell lines producing both mAb and non-mAb biotherapeutics. Glyco-Mapper predicted the altered glycoprofiles and the results were compared against each published glycoform to establish a 96.1% glycan prediction accuracy (1,547 of 1,608 glycans). The average prediction glycoform accuracy, sensitivity, and specificity statistics are 96%, 85%, and 97%, respectively (Table 4.1). The following sections illustrate the application of the Glyco-Mapper tool towards four different examples of glycoform-engineering modification strategies from among those reported in the literature.

Table 4.1: The average accuracy, sensitivity, specificity, and delta accuracy statistics for each cell-engineered glycoform prediction. The average is greater than 80% for all statistics and greater than 95% for both the overall accuracy and specificity.

Author, Year	Gene(s)	Accuracy	Sensitivity	Specificity	Delta Accuracy
Onitsuka, 2012*	*ST6Gal1*	92.5%	83.3%	94.1%	50.0%
Goh, 2013*	*GnT-I / (Fut8)*	93.6%	78.6%	95.1%	81.8%
Kanda, 2007'	*Fut8*	97.5%	100.0%	97.3%	100.0%
Kanda, 2007'	*GMDS*	97.5%	100.0%	97.3%	100.0%
Kanda, 2007*	*GMDS* / Fuc Feed / *(Fut8)*	97.5%	100.0%	97.3%	100.0%
Maszczak, 2013'	*SLC35A3*	98.7%	90.0%	99.3%	83.3%
Maszczak, 2013'	*β4Galt / (GnT-II)*	98.1%	92.3%	98.6%	85.7%
Maszczak, 2013*	*β4Galt / SLC35A3 / (GnT-II)*	98.7%	83.3%	100.0%	90.9%
Kremkow, 2016*	*GnT-II*	94.9%	77.3%	97.8%	75.0%
Malphettes, 2010	*Fut8*	92.5%	85.7%	93.9%	87.5%
Tsukahara, 2004	*Fut8*	97.5%	100.0%	97.3%	100.0%
Naso, 2010	*SiaA*	97.5%	100.0%	97.3%	100.0%
Sealover, 2013	*GnT-I*	92.5%	0.0%	97.4%	100.0%
Imai-Nishiya, 2007	*Fut8 / GMDS*	97.5%	100.0%	97.3%	100.0%
Weikert, 1999	*β4Galt*	95.5%	87.5%	95.9%	0.0%
Weikert, 1999	*ST3Gal3*	94.9%	85.7%	95.3%	0.0%
Weikert, 1999	*β4Galt / ST3Gal3*	96.2%	81.8%	97.2%	-
Average		**96.2%**	**84.6%**	**97.2%**	**84.7%**

4.5.1 Strategy 1: Expression of Heterologous Glycosyltransferases (e.g. – *ST6Gal1*)

Expression of a non-native glycosylation gene may result in the production of novel glycans or shift the glycoform distribution. Glyco-Mapper successfully replicated and predicted the final reference and engineered glycan distributions, respectively, as reported by Naso et al. (2010) and Onitsuka et al. (2012). Onitsuka et al. (2012) expressed *ST6Gal1* to increase sialylation, thereby potentially increasing the IgG's biotherapeutic *in vivo* half-life. Glyco-Mapper replicated the wild type glycoform (Figure 4.1) with 39 of 40 correct glycans [3 of 4 present; 36 of 36 absent]. When *ST6Gal1* expression was estimated (Figure 4.2), Glyco-Mapper predicted 37 of 40 glycans correctly [5 of 7 present; 32 of 33 absent]. The predicted IgG glycoform resulting from the altered heterologous glycosyltransferase flux and expression was accurate, sensitive, and specific.

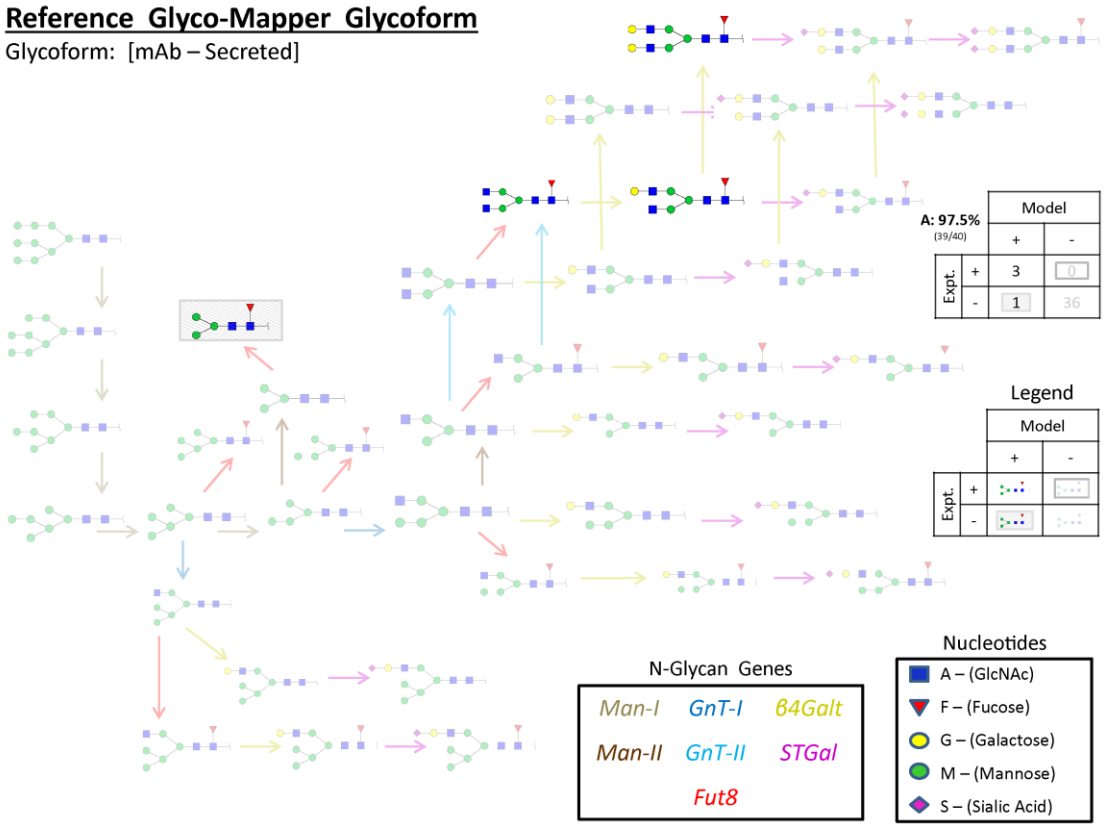


Figure 4.1: The Glyco-Mapper prediction of the Onitsuka *et al.* reference glycoform. The asialylated glycans FA2, FA2G1, and FA2G2 are correctly predicted to be experimentally present.

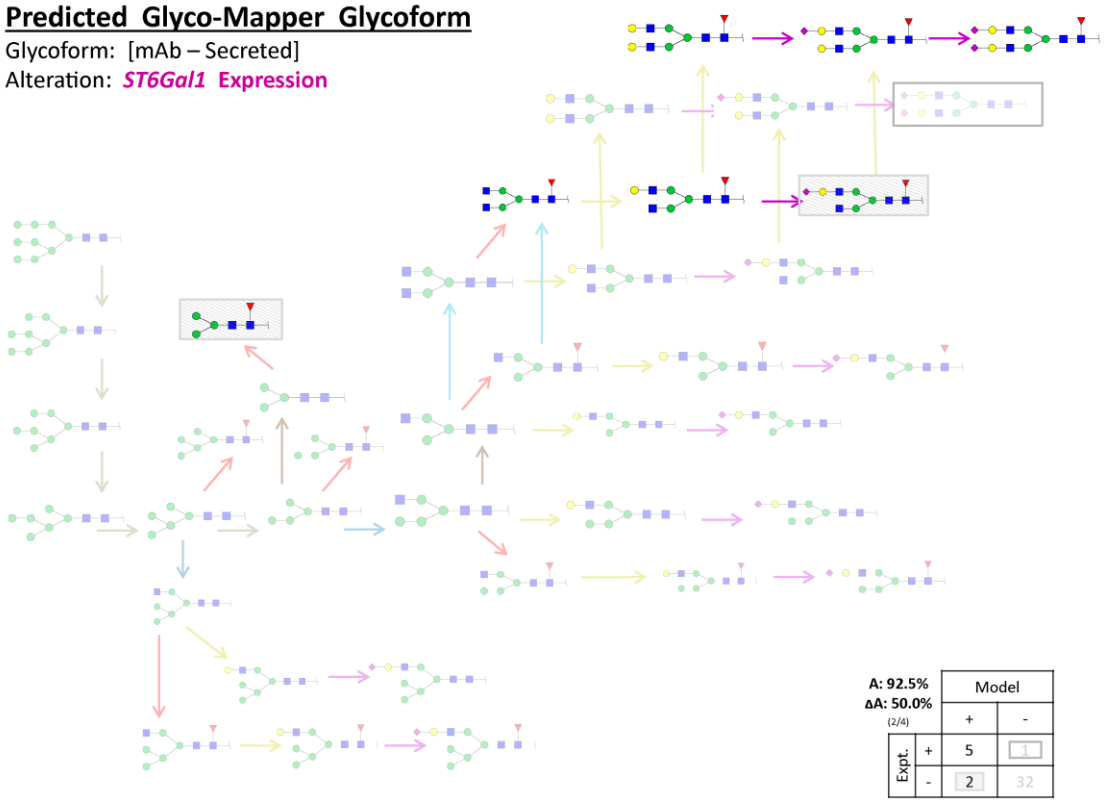


Figure 4.2: The Glyco-Mapper prediction of the expression of *ST6Gal1* based on the Onitsuka *et al.* reference glycoform (Figure 4.1). The sialylated glycans FA2G2S1 and FA2G2S2 are correctly predicted to be experimentally present in this gain-of-function glycoform engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

4.5.2 Strategy 2: Genetic Manipulation of Glycosyltransferases (e.g. – *GnT-I*)

Genome editing tools are increasingly being used to knockdown, knockout, or overexpress targeted glycosylation genes and alter biotherapeutic glycoforms. Glyco-Mapper successfully replicated studies by Kanda et al. (2007), Weikert et al. (1999), Malphettes et al. (2010), Sealover et al. (2013), Goh et al. (2014), Maszczak-Seneczko et al. (2013), and Tsukahara et al. (2006). In particular, Goh et al. (2014) investigated the effect of *GnT-I* expression in a *GnT-I* knockout CHO cell line with the goal of increasing the glycoprotein erythropoietin (EPO) sialylation. Glyco-Mapper accurately replicated 149 of 156 glycans [2 of 5 present; 147 of 151 absent] for the wild type (*GnT-I* knockout) glycoform (Figure 4.3); whereas Glyco-Mapper predicted 146 of 156 glycans correctly [11 of 18 present, 135 of 138 absent] when the *GnT-I* overexpression was estimated (Figure 4.4). The predicted EPO glycoform resulting from the altered glycosyltransferase flux and expression was highly accurate and specific.

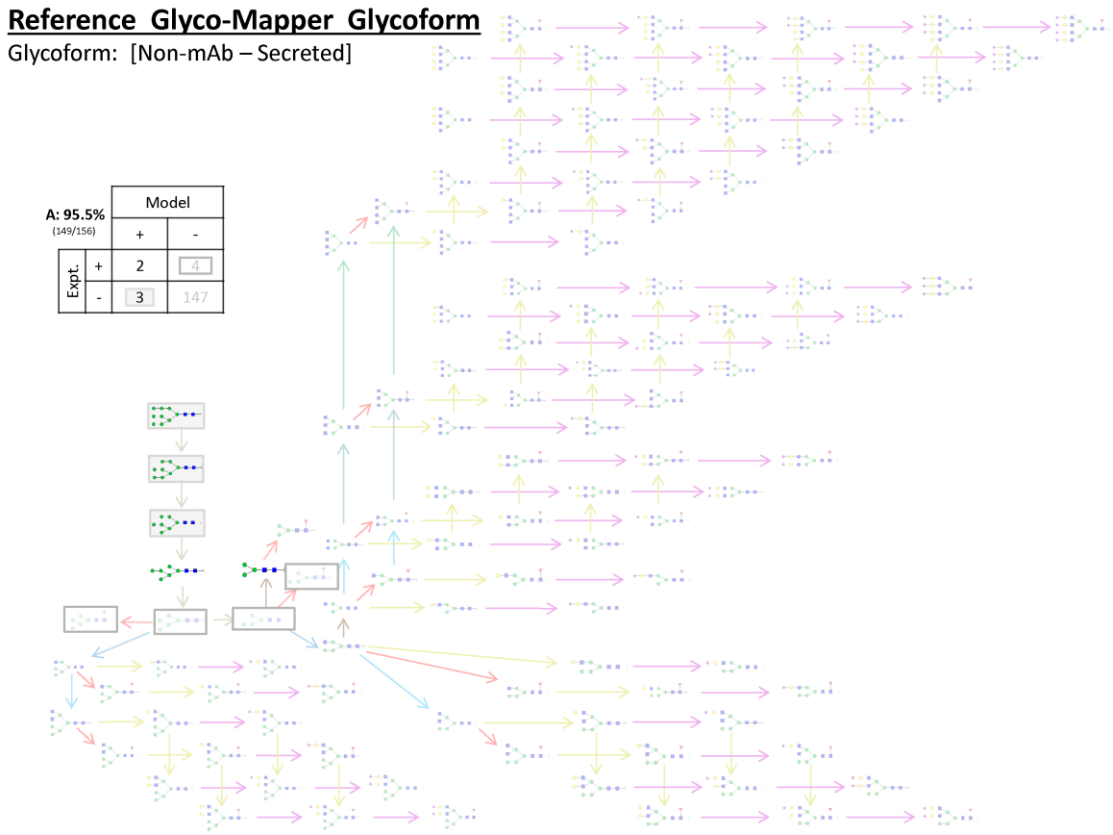


Figure 4.3: The Glyco-Mapper prediction of the Goh *et al.* reference glycoform. The mannose glycans M3 and M6 are correctly predicted to be experimentally present. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

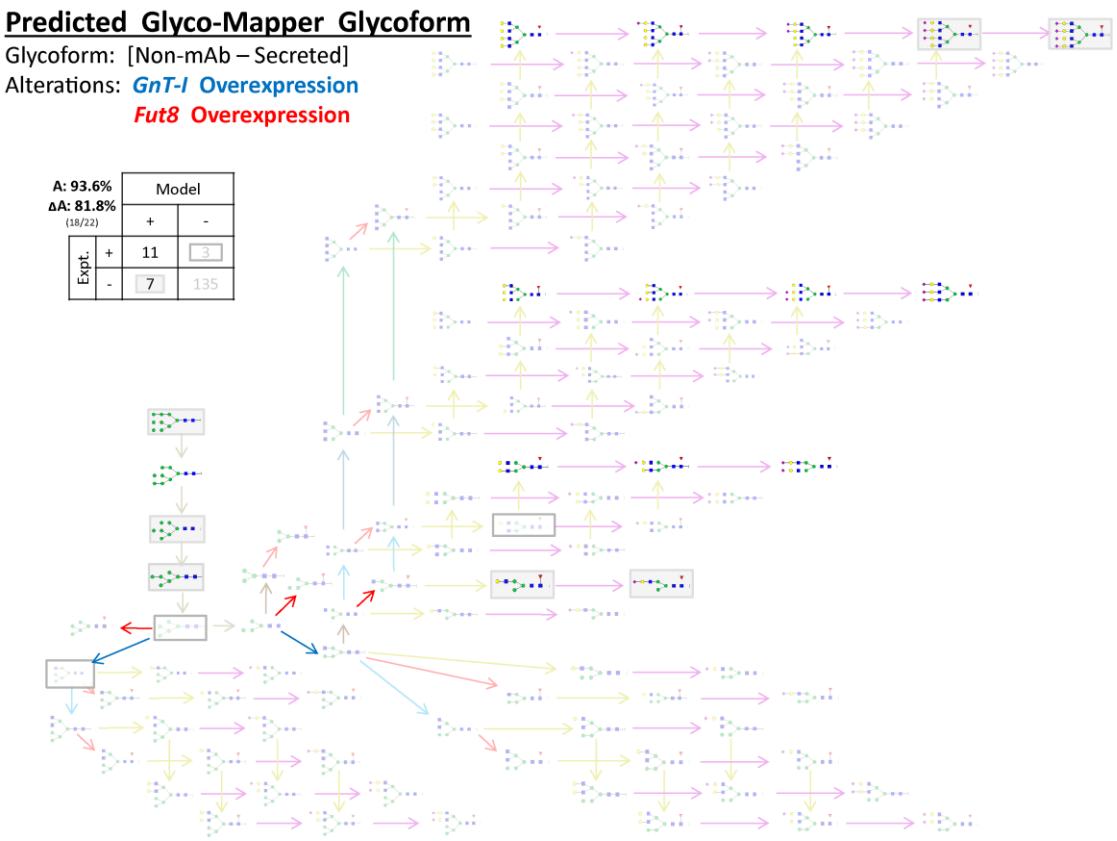


Figure 4.4: The Glyco-Mapper prediction of the *GnT-I* overexpression based on the Goh *et al.* reference glycoform (Figure 4.3). Multiple novel bi-antennary (FA2G2, FA2G2S1, and FA2G2S2), tri-antennary (FA3G3, FA3G3S1, FA3G3S2, and FA3G3S3), and tetra-antennary (FA4G4, FA4G4S1, and FA4G4S2) glycans are all correctly predicted to be experimentally present in this gain-of-function glycoform engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

4.5.3 Strategy 3: Genetic Manipulation of Glycosyltransferase and Metabolism Genes (e.g. – *GMDS* and *Fut8*) and Nutrient Feeding Modifications (e.g. – Fucose Feed)

The knockout of a native metabolism gene or the alteration of a media feed (nutrient composition) can result in a modified glycoform, as reported by Kanda et al. (2007) and Imai-Nishiya et al. (2007), both of which Glyco-Mapper successfully predicted. Using a mAb (IgG1)-producing CHO cell line, Kanda et al. (2007) independently knocked out *GMDS, Fut8*, and *GMDS* with an altered nutrient-feed containing fucose, thereby affecting the fucosylation and antibody-dependent cellular cytotoxicity (Shinkawa et al., 2003). Glyco-Mapper replicated the wild type glycoform (Figure 4.5) with 37 of 40 correct glycans [5 of 8 present, 32 of 32 absent]. When both the *GMDS* knockout (Figure 4.6) and *Fut8* knockout (Figure 4.7) were independently incorporated, Glyco-Mapper accurately predicted 39 of 40 glycans correctly [3 of 4 present; 36 of 36 absent]. Glyco-Mapper incorporated the nutrient-feed containing fucose and the *GMDS* knockout (Figure 4.8) and accurately predicted 39 of 40 glycans correctly [3 of 4 present; 36 of 36 absent]. The predicted IgG1 glycoforms resulting from the modified feeding strategy and the altered metabolic and glycosyltransferase gene fluxes and expressions were all highly accurate, sensitive, and specific.

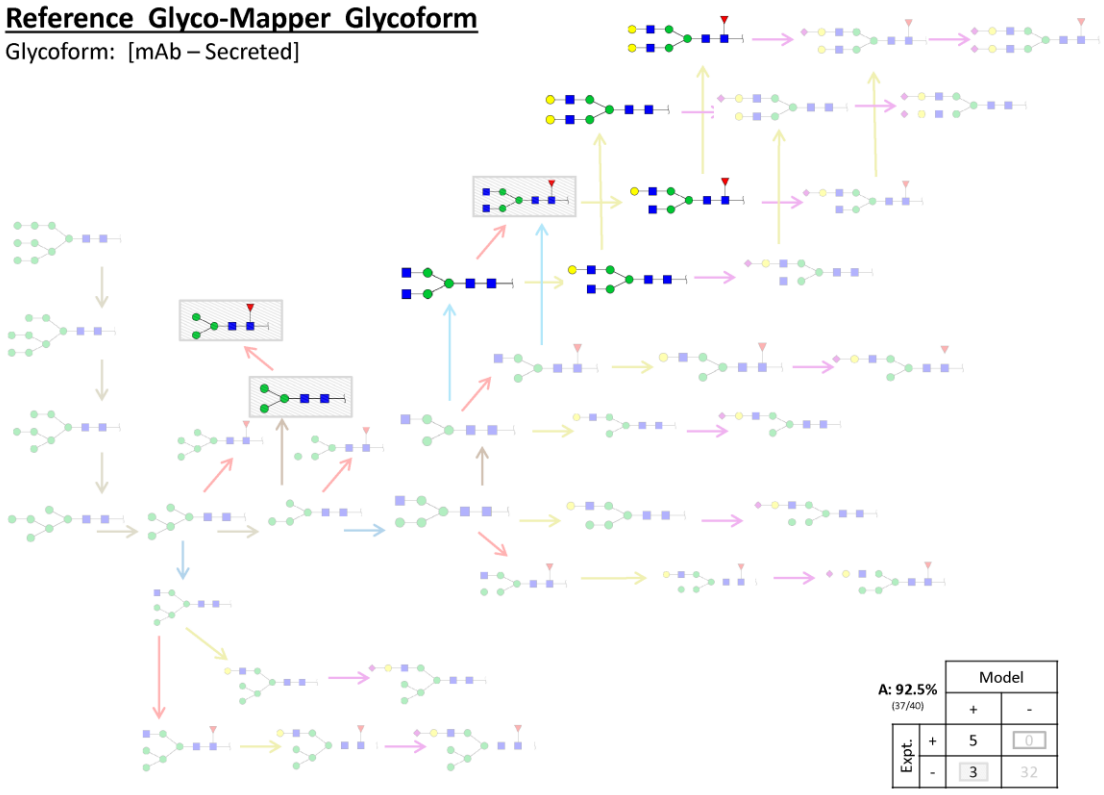


Figure 4.5: The Glyco-Mapper prediction of the Kanda *et al.* reference glycoform. The bi-antennary glycans A2, A2G1, FA2G1, A2G2, and FA2G2 are correctly predicted to be experimentally present. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

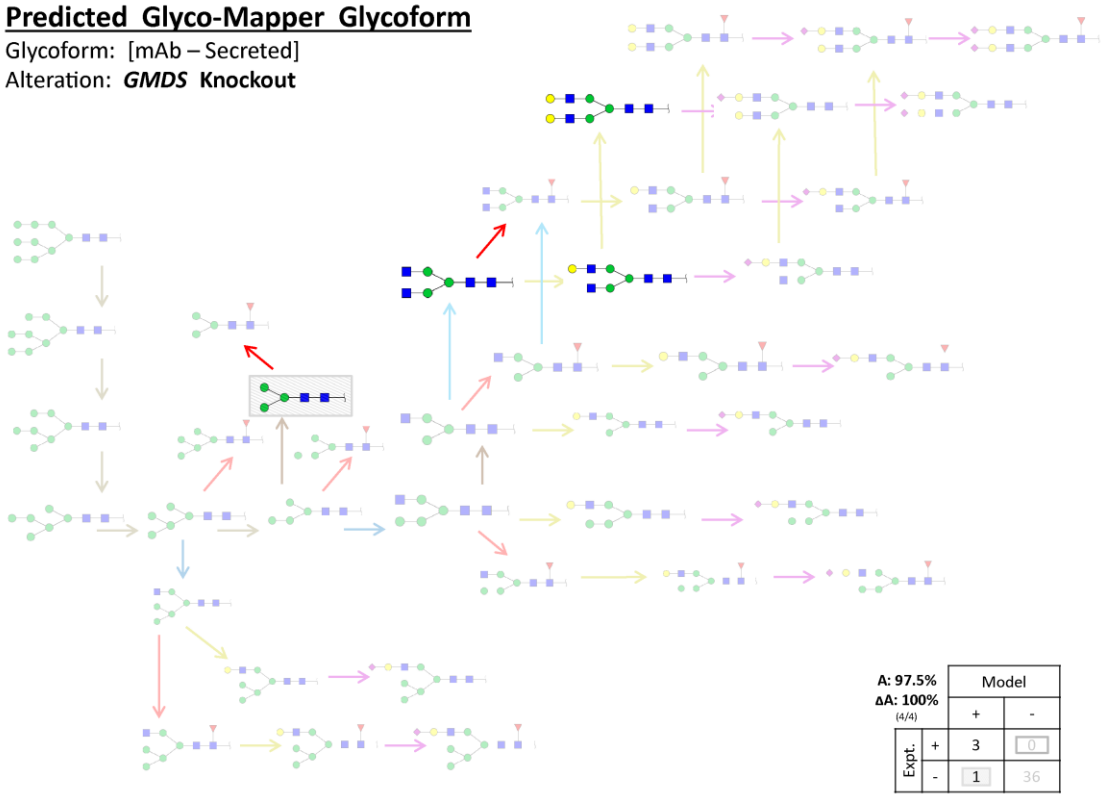


Figure 4.6: The Glyco-Mapper prediction of the knockout of *GMDS* based on the Kanda *et al.* reference glycoform (Figure 4.5). The afucosylated bi-antennary glycans A2, A2G1, and A2G2 are correctly predicted to be experimentally present in this glycoform engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

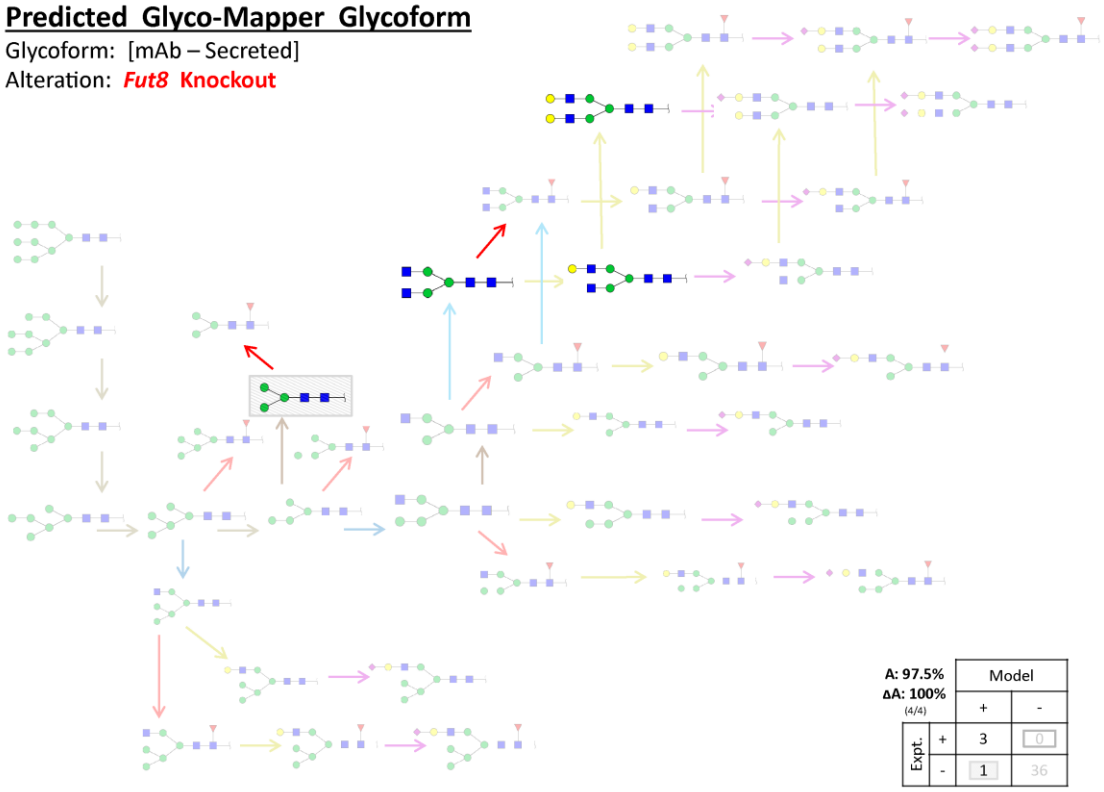


Figure 4.7: The Glyco-Mapper prediction of the knockout of *Fut8* based on the Kanda *et al.* reference glycoform (Figure 4.5). The afucosylated bi-antennary glycans A2, A2G1, and A2G2 are correctly predicted to be experimentally present in this glycoform engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

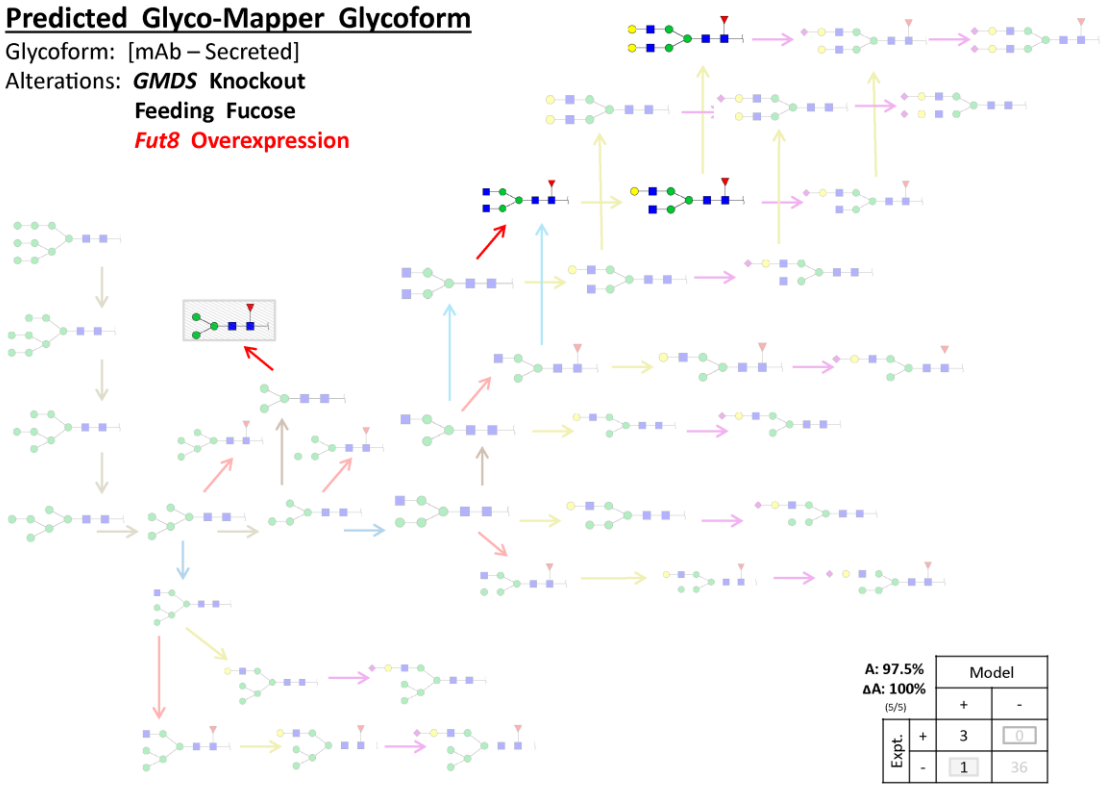


Figure 4.8: The Glyco-Mapper prediction of the fucose feeding strategy coupled with the knockout of *GMDS* based on the Kanda *et al.* reference glycoform (Figure 4.5). The fucosylated bi-antennary glycans FA2, FA2G1, and FA2G2 are correctly predicted to be experimentally present in this metabolic and glycosylation engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

4.5.4 Strategy 4: Genetic Manipulation of Glycosyltransferases and Nucleotide Sugar Transporter Genes (e.g. – *SLC35A3* and *β4Galt*)

Altered nucleotide sugar transport genes affect the glycoform, whether altered independently or in conjunction with a glycosyltransferase. Glyco-Mapper successfully predicted the glycoprotein glycoforms reported by Maszczak-Seneczko et al. (2013) whom knocked out both *SLC35A3*, the gene responsible for UDP-GlcNAc transport and reduced transport results in reduced glycoprotein glycan antennarity, and *β4Galt*, the genes responsible for the addition of Gal nucleotides, which reduce the biologic's half-life when the Gal nucleotides are terminal (Ashwell and Morell, 1974). Glyco-Mapper accurately replicated 153 of 156 glycans [10 of 12 present, 143 of 144 absent] for the wild type glycoform (Figure 4.9); whereas Glyco-Mapper predicted 153 of 156 glycans correctly [12 of 14 present; 141 of 142 absent] when the *β4Galt* knockout was incorporated (Figure 4.10). Glyco-Mapper predicted 154 of 156 glycans correctly [9 of 10 present; 145 of 146 absent] when the *SLC35A3* knockout was incorporated (Figure 4.11). Glyco-Mapper accounted for the combined *SLC35A3* and *β4Galt* knockouts (Figure 4.12) accurately by also predicting 154 of 156 glycans correctly [10 of 10 present; 144 of 146 absent]. The predicted glycoprotein glycoforms resulting from the altered nucleotide sugar transporter and glycosyltransferase gene fluxes and expressions were accurate, sensitive, and specific.

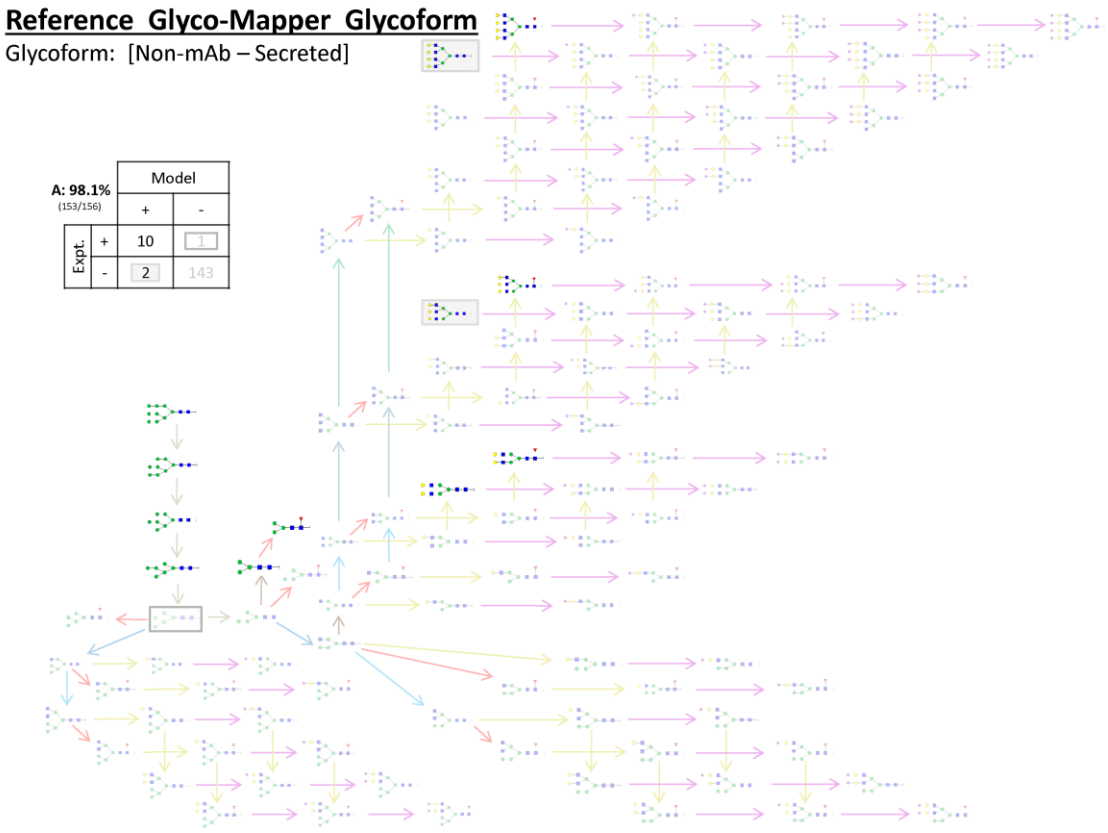


Figure 4.9: The Glyco-Mapper prediction of the Maszczak-Seneczko *et al.* reference glycoform. Bi-, tri-, and tetra-antennary glycans as well as high and low mannose glycans are correctly predicted to be experimentally present. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

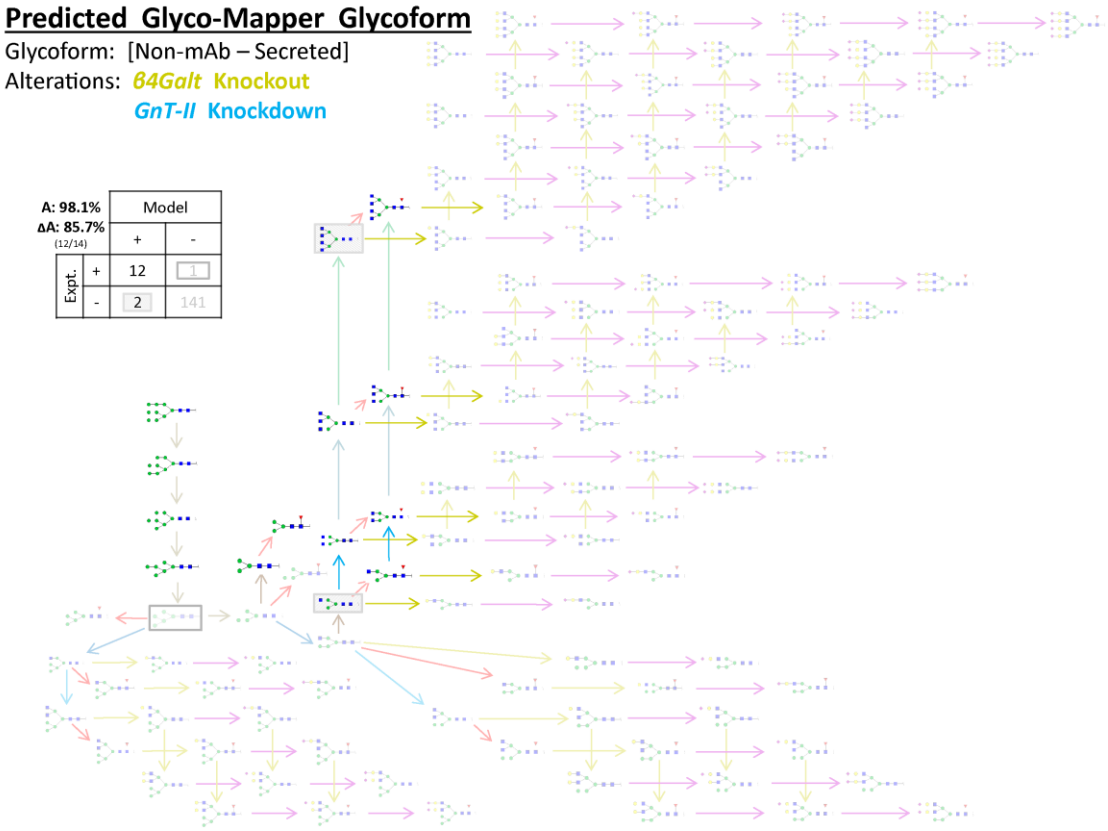


Figure 4.10: The Glyco-Mapper prediction of the *β4Galt* knockout strategy based on the Maszczak-Seneczko *et al.* reference glycoform (Figure 4.9). The agalactosylated glycans FA4, A3, FA3, A2, FA2, and FA1 are all correctly predicted to be experimentally present in this glycoform engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

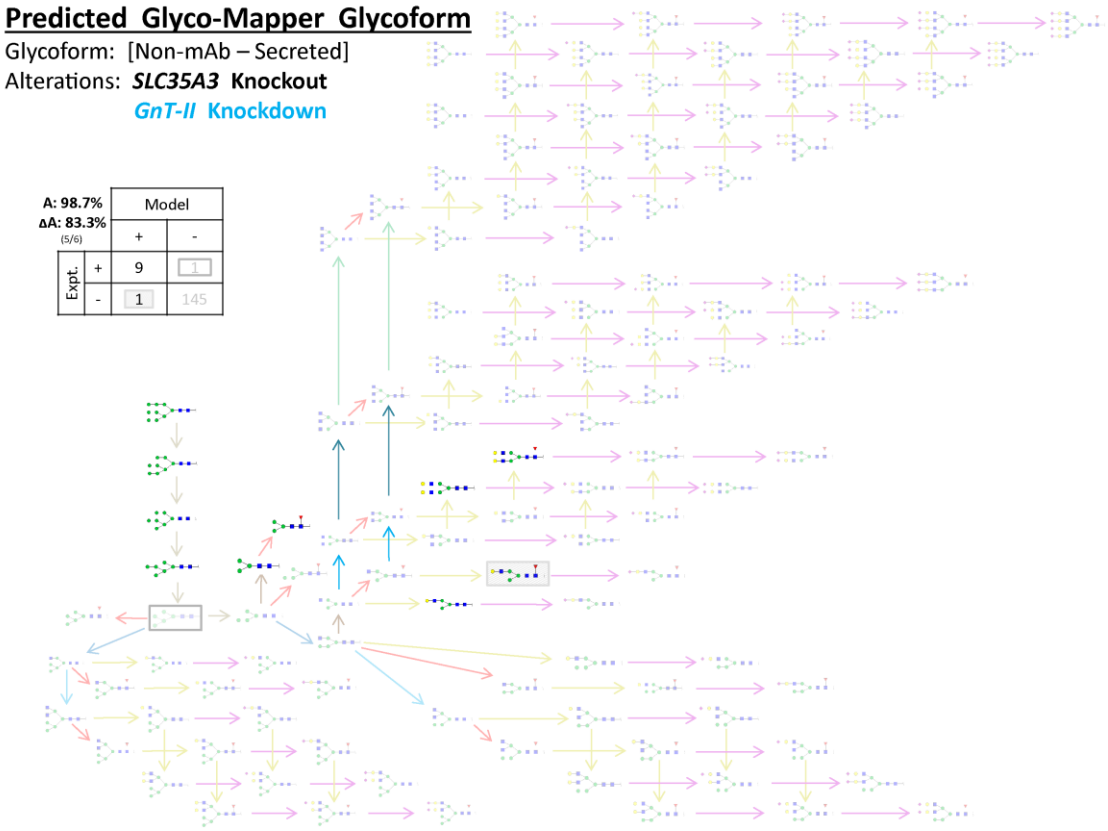


Figure 4.11: The Glyco-Mapper prediction of the *SLC35A3* knockout strategy based on the Maszczak-Seneczko *et al.* reference glycoform (Figure 4.9). The glycans A2G2, FA2G2, and A1G1 are all correctly predicted to be experimentally present in this complex glycosylation engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

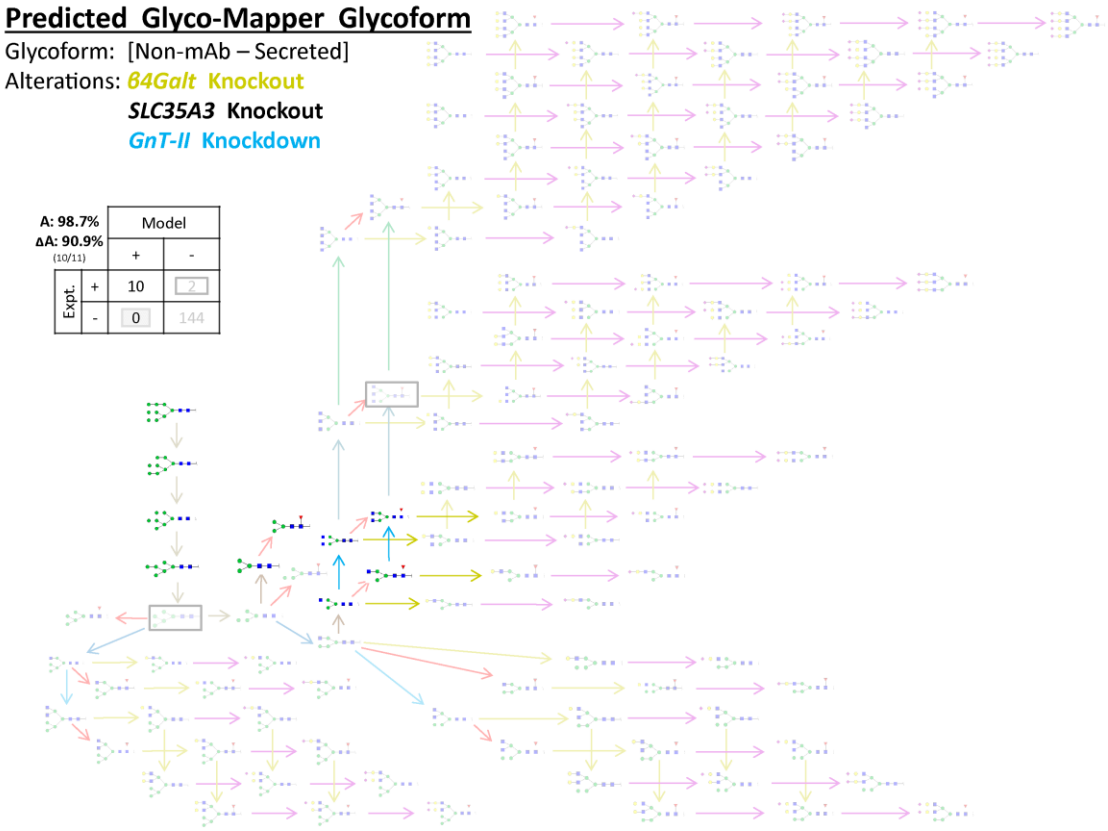


Figure 4.12: The Glyco-Mapper prediction of the *SLC35A3* and *β4Galt* knockout strategy based on the Maszczak-Seneczko *et al.* reference glycoform (Figure 4.9). The agalactosylated glycans A2, FA2, A1, and FA1 are all correctly predicted to be experimentally present in this complex glycosylation engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

4.5.5 Experimental Confirmation of a Novel "Strategy 2" Modification: Genetic Manipulation of Glycosyltransferases (e.g. – *GnT-II*)

After confirmation of Glyco-Mapper's ability to accurately predict reported changes in literature, Glyco-Mapper's ability to predict a non-obvious change not previously defined in literature was experimentally tested. The gene *GnT-II* was knocked down using short interfering RNA (siRNA) with the goal of inhibiting bi-antennary glycan formation of the model glycoprotein secreted alkaline phosphatase (SEAP). Glyco-Mapper accurately replicated 144 of 156 glycans [11 of 14 present, 133 of 142 absent] for the wild type glycoform (Figure 4.13); whereas Glyco-Mapper predicted 148 of 156 glycans correctly [17 of 20 present; 131 of 136 absent] when the *GnT-II* knockdown was estimated (Figure 4.14). The predicted SEAP glycoform resulting from the altered flux and expression of the glycosyltransferase *GnT-II* was novel as well as highly accurate and specific.

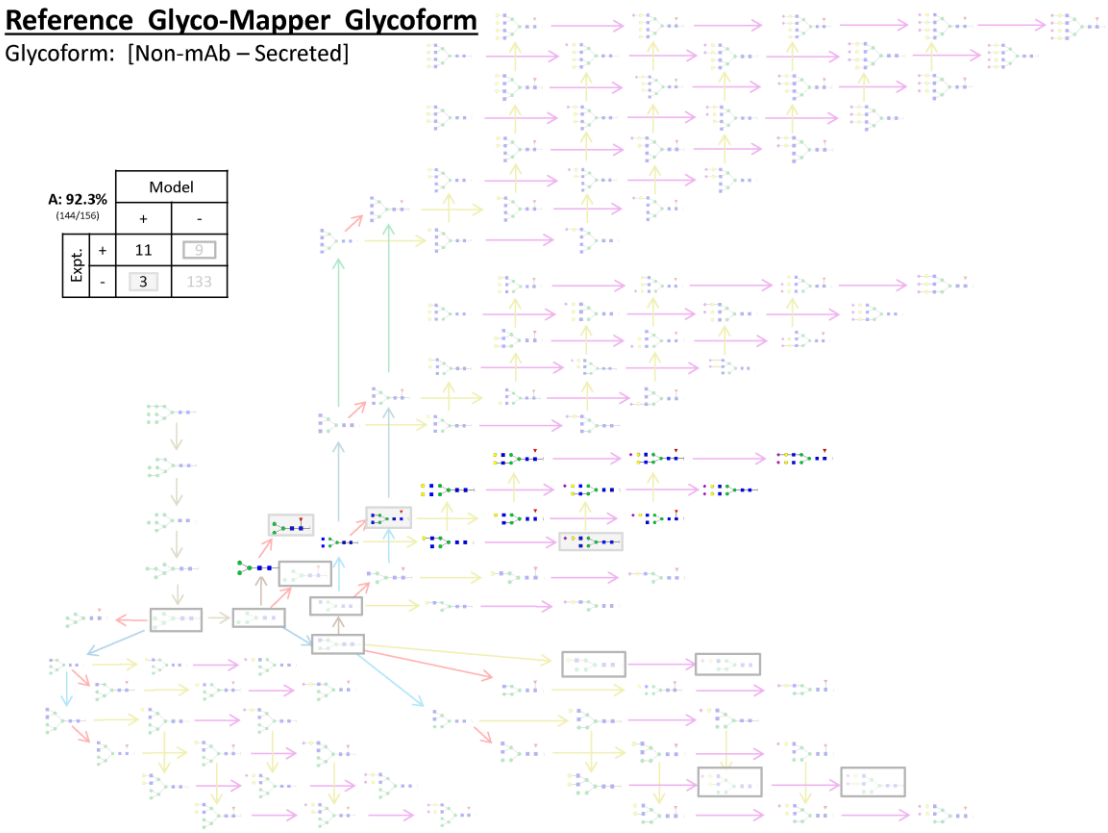


Figure 4.13: The Glyco-Mapper prediction of the reference SEAP glycoform. Most bi-antennary glycans are correctly predicted to be experimentally present. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

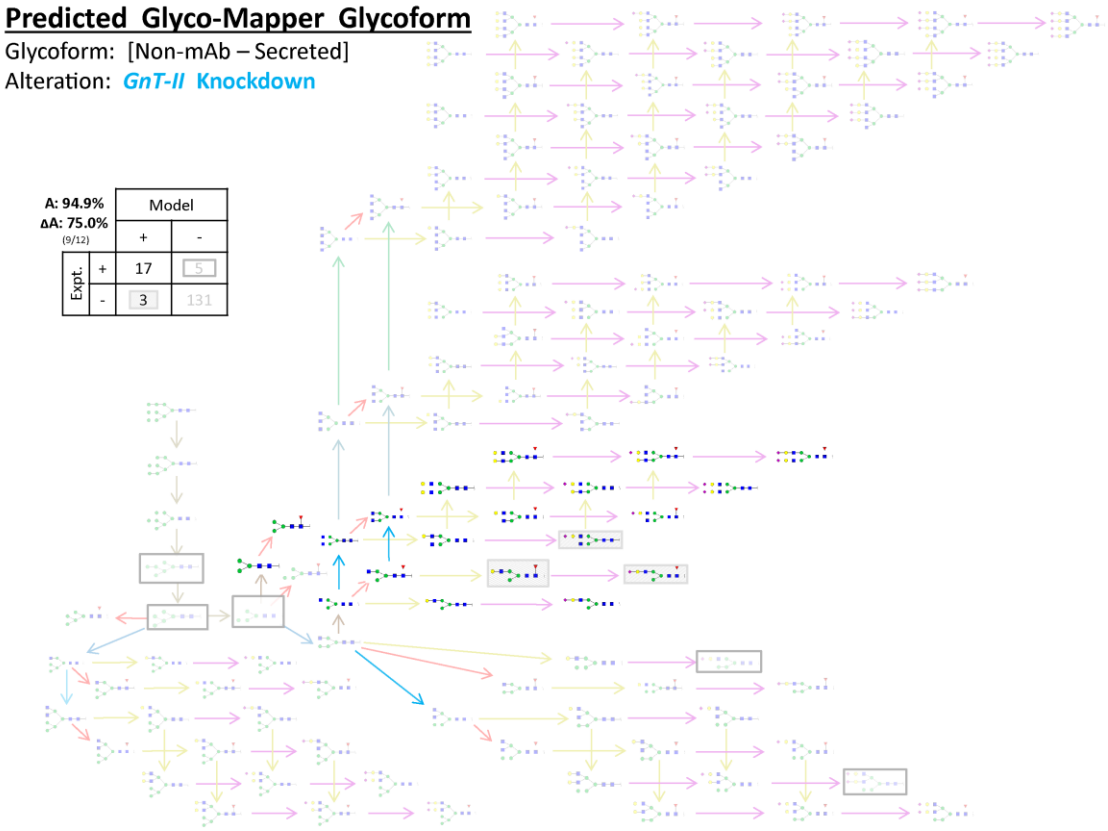


Figure 4.14: The Glyco-Mapper predicted *GnT-II* knockdown glycoform is based on the SEAP reference glycoform (Figure 4.13). The single antennae glycans A1, FA1, A1G1, and A1G1S1 are all correctly predicted to be experimentally present in this glycoform engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure 4.1 are not pictured but still apply.

4.6 Discussion

The Glyco-Mapper tool developed via the DReaM-zyP method builds upon a protocol for the generation of high-quality genome-scale metabolic reconstructions (Thiele and Palsson, 2010) by incorporating discretized parameters representing stoichiometric pathway fluxes to predict glycoforms. Manual reconstruction and refinement of the glycosylation reaction network as well as the central carbon metabolism (CCM) and nucleotide sugar transport metabolic pathways was accomplished using the CHO-K1 and Chinese hamster genome annotations. The incorporation of discrete, quantized reaction flux parameters for each gene in the reaction network and for each media feed sugar transformed the reconstruction database to Glyco-Mapper, a stoichiometric flux modeling tool. Glyco-Mapper provides cell line specific and experimentally relevant glycoform predictions using altered genotypes. The range of glycoform phenotypes accurately predicted by Glyco-Mapper include fucosylation, sialylation, galactosylation, antennarity, nucleotide sugar transport (UDP-GlcNAc), and nucleotide sugar metabolism (GDP-Fuc). Additional details regarding the creation of DReaM-zyP and the use of DReaM-zyP to create Glyco-Mapper are presented in Appendix B.

Glyco-Mapper predicted the overexpression, knockout, and knockdown of glycosylation, nucleotide sugar transporter, and metabolism genes and the predicted glycoforms aligned with the experimental glycoforms with an average sensitivity of 85% and an accuracy and specificity of greater than 95%. The Glyco-Mapper predictions listed in Table 4.1 but not described or pictured in the results section are described and illustrated in Appendix C. Glyco-Mapper also accurately predicted the accuracy of glycans changing experimental classifications (e.g. present to absent)

between the reference and prediction glycoforms with an average accuracy of 85%, referred to here as the delta accuracy. Glycoforms of both mAbs (nine alterations) and non-mAbs (eight alterations) were predicted and Glyco-Mapper predicted both the altered mAb and non-mAb glycoprofiles with an average accuracy, sensitivity, and specificity of 96%, 85%, and 97%, respectively (Tables 4.2 and 4.3). Glyco-Mapper predictions are reliable because the predictive delta accuracy is high and the accuracy, sensitivity, and specificity statistics are consistent for both mAb and non-mAb biotherapeutics.

Table 4.2: The Glyco-Mapper prediction accuracy, sensitivity, specificity, and delta accuracy statistics for mAb biopharmaceuticals.

Author, Year	Gene(s)	Accuracy	Sensitivity	Specificity	Delta Accuracy
Onitsuka, 2012*	*ST6Gal1*	92.5%	83.3%	94.1%	50.0%
Kanda, 2007'	*Fut8*	97.5%	100.0%	97.3%	100.0%
Kanda, 2007'	*GMDS*	97.5%	100.0%	97.3%	100.0%
Kanda, 2007*	*GMDS* / Fuc Feed / *(Fut8)*	97.5%	100.0%	97.3%	100.0%
Malphettes, 2010	*Fut8*	92.5%	85.7%	93.9%	87.5%
Tsukahara, 2004	*Fut8*	97.5%	100.0%	97.3%	100.0%
Naso, 2010	*SiaA*	97.5%	100.0%	97.3%	100.0%
Sealover, 2013	*GnT-I*	92.5%	0.0%	97.4%	100.0%
Imai-Nishiya, 2007	*Fut8 / GMDS*	97.5%	100.0%	97.3%	100.0%
Average		**95.8%**	**85.4%**	**96.6%**	**93.1%**

Table 4.3: The Glyco-Mapper prediction accuracy, sensitivity, specificity, and delta accuracy statistics for non-mAb biopharmaceuticals.

Author, Year	Gene(s)	Accuracy	Sensitivity	Specificity	Delta Accuracy
Goh, 2013*	*GnT-I / (Fut8)*	93.6%	78.6%	95.1%	81.8%
Maszczak, 2013'	*SLC35A3*	98.7%	90.0%	99.3%	83.3%
Maszczak, 2013'	*β4Galt / (GnT-II)*	98.1%	92.3%	98.6%	85.7%
Maszczak, 2013*	*β4Galt / SLC35A3 / (GnT-II)*	98.7%	83.3%	100.0%	90.9%
Kremkow, 2016*	*GnT-II*	94.9%	77.3%	97.8%	75.0%
Weikert, 1999	*β4Galt*	95.5%	87.5%	95.9%	0.0%
Weikert, 1999	*ST3Gal3*	94.9%	85.7%	95.3%	0.0%
Weikert, 1999	*β4Galt / ST3Gal3*	96.2%	81.8%	97.2%	-
Average		**96.3%**	**84.6%**	**97.4%**	**59.5%**

Incorrect glycans are predicted for a multitude of reasons, including biological variance within the glycosylation process; variable glycan detection sensitivity; and substrate, product, enzymatic, or other cellular inhibition mechanisms caused by the genome modification(s) affecting pathway fluxes not accurately portrayed. Glyco-Mapper inaccurately predicts glycans in one of two ways: glycans predicted to be measured that were not experimentally observed and glycans predicted to be absent that were detected. Glyco-Mapper inaccurately predicted 24 of 548 glycans in Figures 4.2, 4.4, 4.8, 4.12, and 4.14 and the 24 glycans are listed in Table 4.4. More than half of these incorrectly predicted glycans are one or two active enzyme reactions away from glycans correctly predicted to be present, making most "incorrect" predictions only slightly off-target. Other errors may be indicative of unidentified inhibitory factors affecting the final glycoform, as opposed to biological variance or inaccurate modeling assumptions. One potential example of an unaccounted factor is the

incorrect A2G2S2 prediction in Figure 4.2. The reference glycoform was fully fucosylated and after *ST6Gal1* overexpression, the predicted glycoform was also fully fucosylated, yet A2G2S2 was reported and composed a significant percentage (~20%) of the experimental glycoform without an explanation.

Table 4.4: The incorrect Glyco-Mapper glycan predictions within Figures 4.2, 4.4, 4.8, 4.12, and 4.14. These glycans span the range of mannose to complex glycan groupings. More than half of these glycans are two or fewer active enzymatic steps removed from an experimentally measured and correctly predicted glycan, demonstrating the complexity of this non-template driven process.

Glycan	Predicted as:	Figure
FM3	Present	4.2
FA2G1S1	Present	4.2
A2G2S2	Absent	4.2
M9	Present	4.4
M7	Present	4.4
M6	Present	4.4
M5	Absent	4.4
M5A1	Absent	4.4
FA1G1	Present	4.4
FA1G1S1	Present	4.4
FA2G1	Absent	4.4
FA4G4S3	Present	4.4
FA4G4S4	Present	4.4
FM3	Present	4.8
M5	Absent	4.12
FA3	Absent	4.12
M6	Absent	4.14
M5	Absent	4.14
M4	Absent	4.14
FA1G1	Present	4.14
FA1G1S1	Present	4.14
A2G1S1	Present	4.14
M4A1G1S1	Absent	4.14
M4A2G2S2	Absent	4.14

4.7 Conclusions

Glyco-Mapper predicts glycoforms with high accuracy, specificity, and selectivity after targeted gene manipulations have occurred by combining the power of genome scale reconstruction with discretized kinetics. Published glycosylation, nucleotide sugar metabolism, and nucleotide sugar transporter gene modifications have all been modeled and the predicted glycoforms averaged a 96% accuracy in specific glycan prediction when compared with experimental results. Upon examination and analysis of the few glycans incorrectly predicted, the majority likely result due to biological variability, experimental error, or an imperfect modeling assumption. Glyco-Mapper facilitates an understanding of the possible and likely effects of altering a gene's activity upon the glycoform, through gene knock-outs, knock-downs, or overexpression. Increasingly accurate predictions enable data-driven selection of beneficial genetic alterations that could be useful to the biopharmaceutical manufacturing community.

4.8 Acknowledgements

I am grateful for support from the National Science Foundation (*1412365, 1247394).* I would like to thank Leila Choe for assistance with mass spectrometry data collection, as well as Dr. Yu-Sung Wu for discussions and guidance regarding protein purification.

REFERENCES

Ahn WS, Antoniewicz MR. (2012) Towards dynamic metabolic flux analysis in CHO cell cultures. Biotechnol J. 7:61-74.

America's Biopharmaceutical Research Companies. (2013) Medicines in development – Biologics – 2013 Report. 1:1-89.

Ashwell G, Morell AG. (1974) Role of surface carbohydrates in hepatic recognition and transport of circulating glycoproteins. Adv Enzymol Relat Areas Mol Biol. 41:99-128.

Bosques CJ, Collins BE, Meador JW, Sarvaiya H, Murphy JL, DelloRusso G, Bulik DA, Hsu IH, Washburn N, Sipsey SF, Myette JR, Raman R, Shriver Z, Sasisekharan R, Venkataraman G. (2010) Chinese hamster ovary cells can produce galactose-α-1,3-galactose antigens on proteins. Nat Biotechnol. 28:1153-1156.

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

Chen N, Koumpouras GC, Polizzi KM, Kontoravdi C. (2012) Genome-based kinetic modeling of cytosolic glucose metabolism in industrially relevant cell lines: Saccharomyces cerevisiae and Chinese hamster ovary cells. Bioprocess Biosyst Eng. 35:1023-1033.

Chowdhury R, Chowdhury A, Maranas CD. (2015) Using gene essentiality and synthetic lethality information to correct yeast and CHO cell genome-scale models. Metabolites. 5:536-570.

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. PNAS. 104:1777-1782.

Goh JSY, Liu Y, Chan KF, Wan C, Teo G, Zhang P, Zhang Y, Song Z. (2014) Producing recombinant therapeutic glycoproteins with enhanced sialylation using CHO-gmt4 glycosylation mutants. Bioengineered. 5:1–5.

Hammond S, Kaplarevic M, Borth N, Betenbaugh MJ, Lee KH. (2012) Chinese hamster genome database: an online resource for the CHO community at www.CHOgenome.org. Biotechnol Bioeng. 109:1353-1356.

Hayduk EJ, Lee KH. (2005) Cytochalasin D can improve heterologous protein productivity in adherent Chinese hamster ovary cells. Biotechnol Bioeng. 90:354-364.

Imai-Nishiya H, Mori K, Inoue M, Wakitani M, Iida S, Shitara K, Satoh M. (2007) Double knockdown of a1,6-fucosyltransferase (FUT8) and GDP-mannose 4,6-dehydratase (GMD) in antibody-producing cells: a new strategy for generating fully non-fucosylated therapeutic antibodies with enhanced ADCC. BMC Biotechnol. 7:84-96.

Jayapal KP, Wlaschin KF, Hu WS, Yap MGS. (2007) Recombinant protein therapeutics from CHO cells - 20 years and counting. Chem Eng Prog. 103:40-47.

Kanda Y, Imai-Nishiya H, Kuni-Kamochi R, Mori K, Inoue M, Kitajima-Miyama K, Okazaki A, Iida S, Shitara K, Satoh M. (2007) Establishment of a GDP-mannose 4,6-dehydratase (GMD) knockout host cell line: A new strategy for generating completely non-fucosylated recombinant therapeutics. J Biotechnol. 130:300–310.

Kanehisa M, Goto S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27-30.

Krambeck FJ, Betenbaugh MJ. (2005) A Mathematical Model of N-Linked Glycosylation. Biotechnol Bioeng. 92:711-728.

Krambeck FJ, Bennum SV, Narang S, Choi S, Yarema KJ, Betenbaugh MJ. (2009) A mathematical model to derive N-glycan structures and cellular enzyme activities from mass spectrometric data. Glycobiology. 19:1163-1175.

Kremkow BG, Baik JY, MacDonald ML, Lee KH. (2015) CHOgenome.org 2.0: Genome resources and website updates. Biotechnol J. 10:931-938.

La Merie Business Intelligence. (2016) Blockbuster biologics 2015. R&D Pipeline News. 10:3-42.

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

Liu G, Neelamegham S. (2014) A Computational Framework for the Automated Construction of Glycosylation Reaction Networks. PLOS ONE. 9:e100939.

Malphettes L, Freyvert Y, Chang J, Liu PQ, Chan E, Miller JC, Zhou Z, Nguyen T, Tsai C, Snowden AW, Collingwood TN, Gregory PD, Cost GJ. (2010) Highly Efficient Deletion of FUT8 in CHO Cell Lines Using Zinc-Finger Nucleases Yields Cells That Produce Completely Nonfucosylated Antibodies. Biotechnol Bioeng. 130:300–310.

Maszczak-Seneczko D, Sosicka P, Olczak T, Jakimowicz P, Majkowski M, Olczak M. (2013) UDP-N-acetylglucosamine Transporter (SLC35A3) Regulates Biosynthesis of Highly Branched N-glycans and Keratan Sulfate. J Biol Chem. 288:21850–21860.

Naso MF, Tam SH, Scallon BJ, Raju TS. (2010) Engineering host cell lines to reduce terminal sialylation of secreted antibodies. mAbs. 2:519-527.

Onitsuka M, Kim WD, Ozaki H, Kawaguchi A, Honda K, Kajiura H, Fujiyama K, Asano R, Kumagai I, Ohtake H, Omasa T. (2012) Enhancement of sialylation on humanized IgG-like bispecific antibody by overexpression of α-2,6-sialyltransferase derived from Chinese hamster ovary cells. Biotechnol Prod Proc Eng. 94:69–80.

Ouyang A, Bennett P, Zhang A, Yang ST. (2007) Affinity chromatographic separation of secreted alkaline phosphatase and glucoamylase using reactive dyes. Process Biochem. 42:561-569.

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

Selvarasu S, Ho YS, Chong WPK, Wong NSC, Yusufi FNK, Lee YY, Yap MGS, Lee DY. (2012) Combined *in silico* modeling and metabolomics analysis to characterize fed-batch CHO cell culture. Biotechnol Bioeng. 109:1415-1429.

Selvarasu S, Karimi IA, Ghim GH, Lee DY. (2010) Genome-scale modeling and *in silico* analysis of mouse cell metabolic network. Mol Biosyst. 6:152-161.

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

Shlomi T, Cabili MN, Herrgård MJ, Palsson BØ, Ruppin E. (2008) Network-based prediction of human tissue-specific metabolism. Nat Biotechnol. 26:1003-1010.

Spahn PN, Hansen AH, Hansen HG, Arnsdorf J, Kildegaard HF, Lewis NE. (2016) A Markov chain model for N-linked protein glycosylation – towards a low parameter tool for model-driven glycoengineering. Metabol Eng. 33:52-66.

Taniguchi N, Honke K, Fukuda M. (2002) Handbook of glycosyltransferases and related genes. Springer, New York, New York, USA.

Thiele I, Palsson BØ. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 5:93-121.

Tsukahara M, Aoki A, Kozono K, Maseki Y, Fukuda Y, Yoshida H, Kobayashi K, Kakitani M, Tomizuka K, Tsumura H. (2006) Targeted disruption of α-1,6-fucosyltransferase (FUT8) gene by homologous recombination in Chinese hamster ovary (CHO) cells. Animal Cell Technol. 14:175-183.

Umaña P, Bailey JE. (1997) A Mathematical Model of N-linked Glycoform Biosynthesis. Biotechnol Bioeng. 55:890-908.

US DHHS FDA. (2015) Quality considerations in demonstrating biosimilarity of a therapeutic protein product to a reference product. 1:1-19.

Walsh G, Jefferis R. (2006) Post-translational modifications in the context of therapeutic proteins. Nat Biotechnol. 24:1241-1252.

Weikert S, Papac D, Briggs J, Cowfer D, Tom S, Gawlitzek M, Lofgren J, Mehta S, Chisholm V, Modi N, Eppler S, Carroll K, Chamow S, Peers D, Berman P, Krummen L. (1999) Engineering Chinese hamster ovary cells to maximize sialic acid content of recombinant glycoproteins. Nat Biotechnol. 17:1116-1121.

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

Chapter 5

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

5.1 Summary of Conclusions

During biopharmaceutical manufacturing, Chinese hamster ovary (CHO) cells produce biopharmaceuticals with a unique glycoform, which must remain constant to ensure product efficacy and patient safety. Predicting these glycoforms with new models and tools could lead to improved biopharmaceutical manufacturing operations and aid biopharmaceutical design development. This dissertation project created the systems biology tool Glyco-Mapper to generate CHO-specific glycoform predictions, which were validated using published glycoform literature. This tool was then applied to predict the glycoforms of various CHO cell engineering strategies that could potentially improve biopharmaceutical manufacturing product quality.

CHO-specific glycosylation enzymes are critical variables that have a direct effect on the biopharmaceutical glycoforms and it is fundamental to fully understand each individual reaction to improve glycosylation modeling's predictive capabilities and accuracy. Correctly incorporating the genes coding for the glycosylation enzymes from the CHO genome was critical. This research systematically explored Sanger and next-generation sequencing methods towards various omics applications and provided an overview of the CHO omics data, especially the CHO and CH genomes, these

methods have generated (Chapter 2). Additional next-generation sequencing methods were explained, a comparison of each technology's sequencing characteristics were examined in relation to the other sequencing technologies, and the application of these technologies towards animal cell research was investigated (Appendix A). The published NCBI RefSeq CHO and CH genomes were made publicly available at CHOgenome.org through a series of data uploads and website improvements to advance the CHO cell community's access to these sequences (Chapter 3).

Using the updated CHOgenome.org website and the available CHO-specific sequences, I developed a database containing every glycosylation and glycosylation-associated metabolism gene within the CHO genome (Tables F.1 and F.2) and manually verified the associated enzymatic characteristics, including the reactants, products, linkages formed or cleaved, and the associated glycosylation classification. The novel technique Discretized Reaction Network Modeling using Fuzzy Parameters (DReaM-zyP) was created using a combination of genome-scale reconstruction, kinetic modeling, and fuzzy logic modeling techniques. DReaM-zyP was applied to create a CHO-specific glycoform prediction tool named Glyco-Mapper (Appendix B). Glyco-Mapper glycoform predictions were validated using cell engineering literature with complete glycoforms published from 1999-2014 (Chapter 4 and Appendix C). The sixteen cell engineered glycoforms were created using four different cell engineering glycoform alteration strategies and the Glyco-Mapper predictions had an average accuracy of 95%. Optimized experimental methods were applied to explore a novel Glyco-Mapper prediction regarding the *GnT-II* knockdown effects on a non-mAb model biopharmaceutical and the Glyco-Mapper prediction was validated

(Figure 4.14) with 94.9% accuracy and 75% delta accuracy (9 of 12 glycans changed classifications as predicted).

Of the hundreds of glycans CHO cells create, tailoring glycosylation models to predict cell line-specific glycoforms will enhance the ability of biopharmaceutical engineers to understand and control the glycoforms relevant to their specific biopharmaceutical. This specificity will consequently enable the further investigation of precise glycoform engineering controllability during biopharmaceutical manufacturing. The work presented here identifies the effects of CHO-specific gene altering and nutrient feeding glycoform engineering strategies on biopharmaceutical manufacturing and provides a foundation for additional opportunities to explore modeling and predicting industrial biopharmaceutical glycoforms.

5.2 Future Work

The creation of the Glyco-Mapper tool and the resulting glycoform predictions presented in this dissertation provide a foundation for additional CHO-specific glycoform exploration. Extension of this work could improve glycoform controllability and enhance biopharmaceutical product quality. Potential applications include: (1) further developing the Glyco-Mapper tool, (2) validating additional novel non-mAb Glyco-Mapper predictions, and (3) exploring industrially-relevant biopharmaceutical glycoforms and the causes of erroneous predictions.

5.2.1 Glyco-Mapper Modeling Improvements

The Glyco-Mapper modeling and prediction of CHO glycoforms presented in Chapter 4 could be enhanced by increasing the biological modeling accuracy of the incorporated variables. The activity level of each enzyme is represented by the discretized k_{ALV} parameter, but the fuzzification scheme, the scale for converting biological activity values to discretized parameters (Sokhansanj et al. 2009), has not been established. Defining a fuzzification scheme to generate the reference set of k_{ALV} parameters from reference enzyme activity levels would enable direct parameter estimation from the experimental data, if available. The enzyme localization of Glyco-Mapper glycosylation, metabolism, nucleotide sugar production, and nucleotide sugar transport enzymes is only innately incorporated. Explicit incorporation of the glycosylation enzyme cellular localization distributions within the ER and Golgi could be applied to enable a more accurate representation of the recent localization research (Ferrari et al. 2012; van Dijk et al. 2008). Explicit localization of the metabolism, nucleotide sugar production, and nucleotide sugar transport enzymes would likely not affect the glycoform predictions as the current innate localization accounts for separate organelles. However, if additional competing reactions are included in the future, explicit localization would depict a more accurate representation of the biological reality, thereby improving the accuracy of the reaction network and resulting predicted glycoforms.

The Glyco-Mapper glycoform prediction tool could also be improved by incorporating additional glycosylation variables. The addition of a general energy (e.g. ATP) balance check to confirm the energy required by the glycosylation pathway is produced by the CCM reaction network would increase the effect of k_{ALV} parameter

adjustments for reactions 8-24 (Figure F.1) upon the glycosylation reaction network. Incorporation of the availability of metal ions and phosphate groups that are defined in literature (Taniguchi et al. 2002) and are required for the glycosyltransferase enzymatic reactions, but are not currently considered within the Glyco-Mapper code, would further increase the number of biologically important variables. Accounting for the phosphate groups required for sugar nucleotide synthesis would warrant the integration of the phosphate group reaction network enzymes to the database. Cellular energy, metal ions, and phosphate groups are required for glycosylation and their inclusion in the Glyco-Mapper tool would result in an increase in the model's biological accuracy. However, all of these variables are also used for other cellular functions; therefore, their accuracy and the degree of Glyco-Mapper glycoform controllability would greatly depend upon the cellular biology and how the variables are integrated within the Glyco-Mapper.

Glyco-Mapper's predicted glycoforms could produce numerical glycan distributions similar to kinetic model simulations by incorporating a defuzzification scheme as well as additional cellular conditions, such as physical cellular conditions and the cellular growth rate. A defuzzification scheme (Sokhansanj et al. 2009) could produce a quantified glycan composition output in addition to the discretized glycoform, which may be of greater use to industrial Glyco-Mapper applications. Accounting for physical cellular components such as the pH, temperature, osmolality, and other reactor conditions (Ahn et al. 2008; Trummer et al. 2006; Rivinoja et al. 2009; Axelsson et al. 2001) would enable the Glyco-Mapper to capture the effects of macro-scale variables upon the biotherapeutic product quality in addition to the micro-scale variables. Incorporating the cellular growth rate, another variable cited for direct

effect upon glycoform distributions (Hossler et al. 2009), should also be explored. While these improvements may be more difficult to integrate and will likely change the scope of the Glyco-Mapper, this cost should be weighed against the increased biological accuracy of the model and, at the very least, be thoroughly investigated.

The Glyco-Mapper glycosylation database contains only the N-glycosylation gene classification subset within the CHO genome. As O-glycosylation is biopharmaceutically relevant (Higuchi et al. 1992), creation of a Glyco-Mapper O-glycosylation database may also prove worthwhile. The original gene database I created contained every CHO glycosylation gene and each gene's associated information, specifically the gene sequence, reaction reactants, products, linkages formed or cleaved, and the associated glycosylation classification. The genes were then all separated into their respective classifications and only the N-glycosylation classification was imported into the Glyco-Mapper. However, the O-glycosylation classification genes have already been manually curated and organized in a separate spreadsheet. The metabolism portion of Glyco-Mapper was designed to include all possible nucleotide sugars relevant to glycosylation, thus the required work is to import the gathered relevant gene information into Glyco-Mapper, unless additional updates are desired. While mAbs are not commonly O-glycosylated, having a predictive O-glycoform model may prove of worth as some biopharmaceuticals (erythropoietin) do undergo O-glycosylation (Higuchi et al. 1992).

5.2.2 Experimental Validation of Additional Glyco-Mapper Predictions

Targeted glycoform alterations are achieved through cell engineering strategies, such as gene knock-outs, knock-downs, or overexpression, with a variety of

different techniques including clustered regularly interspaced short palindromic repeats (CRISPR), RNA interference (RNAi), and plasmid transfection or lipofectamine technologies. In the knock-out and knock-down methods, nucleic material affects the target gene sequence, yet each method alters the native production level of the gene sequence. The well-characterized reference SEAP glycoform is illustrated in Figure 5.1 and is roughly 50% fucosylated, 70% bi-antennary, and 70% sialylated. This glycoform permits a variety of the compelling cell engineering strategies to be experimentally observed. A few of these SEAP glycoform predictions are summarized in Table 5.1, where the numbers indicate the number of glycans predicted to change classifications from Figure 5.1. The predicted *GnT-II* knockdown is the prediction that was confirmed experimentally in Chapter 4 (Figure 4.14) with nearly 95% accuracy. Any number of the remaining predictions should be experimentally performed as further confirmation of the Glyco-Mapper glycoform predictions.

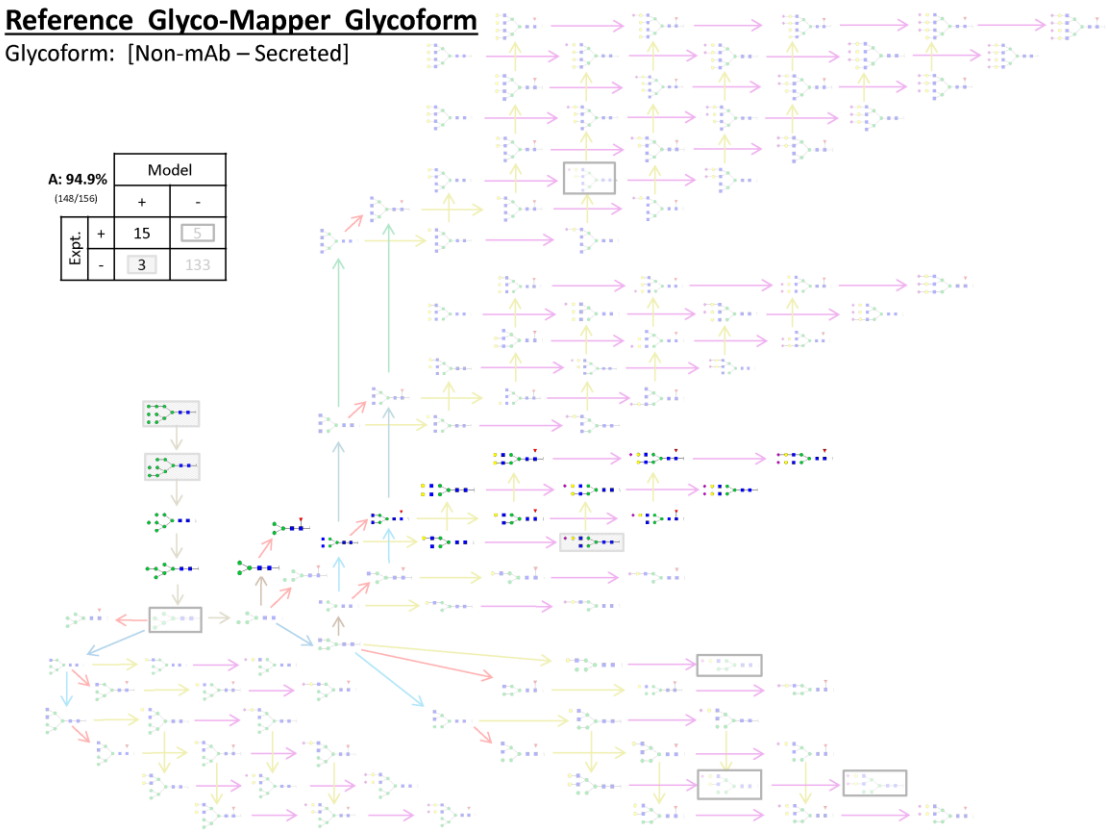


Figure 5.1: CHO-SEAP reference glycoform and the Glyco-Mapper model of the reference glycoform.

Table 5.1: Predicted Glyco-Mapper glycans classified differently for the CHO-SEAP gene alterations compared to the reference glycoform (Figure 5.1). Within the table, the number to the left of the slash represents the number of glycans predicted to be present while the number to the right of the slash are number predicted to be absent. The black numbers represent the single alteration while the blue numbers represent a combination of gene alterations. The gray box identifies the *GnT-II* knockdown shown in Figure 4.14 while the tan boxes identify the mentioned *Fuca1* and *GnT-IV* overexpressions.

Gene		Man2A1/2/B1		**GnT-II**	**GnT-IV**	Fut8		**Fuca1**	Glb1	ST3Gal3		Neu1/2	β4Galnt3/4	β4Galt	
	Effect	KD	OE	**KD**	**OE**	KD	OE	**OE**	OE	KD	OE	OE	OE	KD	OE
β4Galt	OE	11/6	4/6	4/6	8/6-18	0/10	0/10	0/10	0/6	0/10	0/6	0/10	4/6	-	0/6
	KD	11/6	4/6	4/6	8/6-18	0/10	0/10	0/10	0/6	0/10	0/6	0/10	4/6	0/6	-
β4Galnt3/4	OE	19/0	12/0	12/0	4/0-12	4/7	2/7	4/7	4/6	4/6	4/6	4/6	8/0	-	-
Neu1/2	OE	11/6	4/6	4/6	8/6-18	0/10	0/10	0/10	0/10	0/6	0/6	0/6	-	-	-
ST3Gal3	OE	11/6	4/6	4/6	8/6-18	0/10	0/10	0/10	0/6	-	0/6	-	-	-	-
	KD	11/6	4/6	4/6	8/6-18	0/10	0/10	0/10	0/10	0/6	-	-	-	-	-
Glb1	OE	11/6	4/6	4/6	8/6-18	0/10	0/10	0/10	0/6	-	-	-	-	-	-
Fuca1	**OE**	7/7	3/7	3/7	10/7-19	0/7	0/0-14	**0/7**	-	-	-	-	-	-	-
Fut8	OE	7/7	3/7	3/7	10/7-19	-	0/7	-	-	-	-	-	-	-	-
	KD	15/0	6/0	6/0	10/7-19	0/7	-	-	-	-	-	-	-	-	-
GnT-IV	**OE**	59/0-24	26/0-12	26/0-12	**20/0(12)**	-	-	-	-	-	-	-	-	-	-
GnT-II	**KD**	21/0	6/0	**6/0**	-	-	-	-	-	-	-	-	-	-	-
Man2A1/2/B1	OE	-	6/0	-	-	-	-	-	-	-	-	-	-	-	-
	KD	15/0	-	-	-	-	-	-	-	-	-	-	-	-	-

[KD=knock-down; OE=overexpression]

Of the sixteen engineered glycoforms reported in literature (Chapter 4 and Appendix C), eleven used gene knock-outs or knock-downs and five used homologous or heterologous gene overexpression. Despite this decreased frequency of published gene overexpression studies, there are many interesting gene overexpressions to investigate. Two overexpressions depicted in Table 5.1, the overexpression of *GnT-IV* and the overexpression of *Fuca1*, are of particular interest. The overexpression of *GnT-IV*, while not novel with regards to prior publication of CHO cell engineering (Yin et al. 2015), has not been performed on a primarily bi-antennary glycoform of a non-mAb biologic, to the best of my knowledge. The function of GnT-IV is to attach a GlcNAc nucleotide to the core Man (α-1,3) nucleotide with an β-1,4 linkage (Brooks et al. 2002). Determining the increase of tri-antennarity that is achievable should be

explored as altered antennarity has been correlated with altered cell proliferation and differentiation (Lau et al. 2007). The overexpression of *Fuca1* also warrants investigation as *Fuca1* overexpression in any CHO cell has not previously been published to the best of my knowledge. The function of Fuca1 is to trim the terminal Gal nucleotides from the glycan (Taniguchi et al. 2002). As terminal Gal nucleotides decrease the biopharmaceutical half-life (Ashwell and Morell 1974) and more than 30% of the SEAP glycoform has at least one terminal Gal nucleotide, the overexpression of *Fuca1* would be experimentally observable and of potential medical importance. Despite only the overexpression of *Fuca1* being truly novel within CHO in regards to publication, further investigation of both *Fuca1* and *GnT-IV* overexpression is warranted as both genes affect the biopharmaceutical product quality impact on the patient and could further evaluate Glyco-Mapper predictions using the current experimental system.

5.2.3 Additional Exploration of Industrially-Relevant Biopharmaceutical Glycoforms

This work modeled eleven different published reference glycoforms using Glyco-Mapper, of which the majority (seven) were mAbs. The predicted CHO-SEAP glycoform differences in Table 5.1 should be experimentally pursued, but Glyco-Mapper predictions for an industrial mAb reference glycoform should also be conducted. Non-mAb predictions were generated and experimentally pursued in this work because there is a much larger glycan variety within non-mAb glycoforms, making their predictions increasingly difficult compared to mAbs. However, as more than half of the CHO biopharmaceuticals on the market are mAbs (La Merie 2016),

industrial mAb glycoform predictions should also be generated and experimentally validated.

Glycoform variations within biopharmaceuticals are currently difficult to control because of the limited predictive capacity of current models and an incomplete understanding of the interconnectivity of the variables affecting the glycosylation reaction network. The identities of the glycans most often incorrectly predicted should be used to identify and direct research aims. In regards to this work, incorrect Glyco-Mapper predictions should be thoroughly investigated to identify and appropriately adjust the model code. However, not all predictions indicate a coding error, as these errors may identify unknown enzyme or cellular variable interactions or reaction requirements. The incorrect prediction of the glycan A2G2S2 in Figure 4.2 is likely one instance where the incorrect glycan prediction indicates a correlation not currently fully understood. This incorrect prediction was explained in Chapter 4 as "where *ST6Gal1* overexpression caused the unfucosylated glycan A2G2S2, which was absent (along with all other afucosylated glycans) before the overexpression, to compose a significant percentage (~20%) of the experimental glycoform." Multiple publications have reported instances where either the sialylation or fucosylation of a biopharmaceutical was altered and both the sialylation and fucosylation compositions were unexpectedly modified as a result (Onitsuka et al. 2012; Nam et al. 2008). As this correlation does not align with the current glycosylation reaction network understanding, the relationship between fucosylation and sialylation should be examined. Both fucosyaltion and sialylation have critical biopharmaceutical effects on patients and the SEAP glycoform is both partially fucosylated and sialylated, so a targeted increase in fucosylation could be engineered to determine the effect on the

sialylation and vice versa. The experimental results may help define the correlation between changes in sialylation and fucosylation, both of which are vital biopharmaceutical glycosylation components.

REFERENCES

Ahn WS, Jeon JJ, Jeong YR, See SJ, Yoon SK. (2008) Effect of culture temperature on erythropoietin production and glycosylation in a perfusion culture of recombinant CHO cells. Biotechnol Bioeng. 101:1234-1244.

Ashwell G, Morell AG. (1974) Role of surface carbohydrates in hepatic recognition and transport of circulating glycoproteins. Adv Enzymol Relat Areas Mol Biol. 41:99-128.

Axelsson MAB, Karlsson NG, Steel DM, Ouwendijk J, Nilsson T, Hansson GC. (2001) Neutralization of pH in the Golgi apparatus causes redistribution of glycosyltransferases and changes in the O-glycosylation of mucins. Glycobiology. 11:633-644.

Brooks SA, Dwek MV, Schumacher U. (2002) Functional and molecular glycobiology. BIOS Scientific Publishers Limited, Oxford, UK.

Ferrari ML, Gomez GA, Maccioni HJF. (2012) Spatial organization and stoichiometry of n-terminal domain-mediated glycosyltransferase complexes in Golgi membranes determined by fret microscopy. Neurochem Res. 37:1325-1334.

Higuchi M, Oheda M, Kuboniwa H, Tomonoh K, Shimonaka Y, Ochi N. (1992) Role of sugar chains in the expression of the biological-activity of human erythropoietin. J Biol Chem. 267:7703-7709.

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

La Merie Business Intelligence. (2016) Blockbuster biologics 2015. R&D Pipeline News. 10:3-42.

Lau KS, Partridge EA, Grigorian A, Silvescu CI, Reinhold VN, Demetriou M, Dennis JW. (2007) Complex N-glycan number and degree of branching cooperate to regulate cell proliferation and differentiation. Cell. 129:123-134.

Nam JH, Zhang F, Ermonval M, Linhardt RJ, Sharfstein ST. (2008) The effects of culture conditions on the glycosylation of secreted human placental alkaline phosphatase produced in Chinese hamster ovary cells. Biotechnol Bioeng. 100:1178-1192.

Onitsuka M, Kim WD, Ozaki H, Kawaguchi A, Honda K, Kajiura H, Fujiyama K, Asano R, Kumagai I, Ohtake H, Omasa T. (2012) Enhancement of sialylation on humanized IgG-like bispecific antibody by overexpression of α-2,6-sialyltransferase derived from Chinese hamster ovary cells. Biotechnol Prod Proc Eng. 94:69–80.

Rivinoja A, Hassinen A, Kokkonen N, Kauppila A, Kellokumpu S. (2009) Elevated Golgi pH impairs terminal N-glycosylation by inducing mislocalization of Golgi glycosyltransferases. J Cell Physiol. 220:144-154.

Sokhansanj BA, Datta S, Hu X. (2009) Scalable dynamic fuzzy biomolecular network models for large scale biology. Fuzzy Systems in Bio. 1:235-255.

Taniguchi N, Honke K, Fukuda M. (2002) Handbook of glycosyltransferases and related genes. Springer, New York, New York, USA.

Trummer E, Fauland K, Seidinger S, Schriebl, Lattenmayer C, Kunert R, Vorauer-Uhl K, Weik R, Borth N, Katinger H, Muller D. (2006) Process parameter shifting part I. Effect of DOT, pH and temperature on the performance of Epo-Fc expressing CHO cells cultivated in controlled batch bioreactors. Biotechnol Bioeng. 94:1033-1044.

Van Dijk ADJ, Bosch D, ter Braak CJF, van der Krol AR, van Ham RCHJ. (2008) Predicting sub-Golgi localization of type II membrane proteins. Bioinformatics. 24:1779-1786.

Yin B, Gao Y, Chung CY, Yang S, Blake E, Stuczynski MC, Tang J, Kildegaard HF, Anderson MR, Zhang H, Betenbaugh MJ. (2015) Glycoengineering of Chinese hamster ovary cells for enhanced erythropoietin N-glycan branching and sialylation. Biotechnol Bioeng. 112:2343-2351.

REFERENCES

Ahn WS, Antoniewicz MR. (2012) Towards dynamic metabolic flux analysis in CHO cell cultures. Biotechnol J. 7:61-74.

Ahn WS, Jeon JJ, Jeong YR, See SJ, Yoon SK. (2008) Effect of culture temperature on erythropoietin production and glycosylation in a perfusion culture of recombinant CHO cells. Biotechnol Bioeng. 101:1234-1244.

Allay Jr. WR, Mann BF, Novotny MV. (2013) High-sensitivity analytical approaches for the structural characterization of glycoproteins. Chem Rev. 113: 2668-2732.

America's Biopharmaceutical Research Companies. (2013) Medicines in development – Biologics – 2013 Report. 1:1-89.

Ansorge WJ. (2009) Next-generation DNA sequencing techniques. New Biotechnol. 25:195-203.

Ashwell G, Morell AG. (1974) Role of surface carbohydrates in hepatic recognition and transport of circulating glycoproteins. Adv Enzymol Relat Areas Mol Biol. 41:99-128.

Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA, Bakel Hv, Schadt EE, Reijo-Pera RA Underwood JG, Wong WH. (2013) Characterization of the human ESC transcriptome by hybrid sequencing. P Natl Acad Sci USA. 110:E4821-E4830.

Axelsson MAB, Karlsson NG, Steel DM, Ouwendijk J, Nilsson T, Hansson GC. (2001) Neutralization of pH in the Golgi apparatus causes redistribution of glycosyltransferases and changes in the O-glycosylation of mucins. Glycobiology. 11:633-644.

Baik JY, Lee KH. (2014) miRNA expression in CHO:Nature knows best. Biotechnol J. 9:459-460.

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

Baker M. (2010) Nanotechnology imaging probes: smaller and more stable. Nat Methods. 7:957-962.

Bartel DP. (2009) MicroRNAs: target recognition and regulatory functions. Cell. 136:215-233.

Baycin-Hizal D, Tabb DL, Chaerkady R, Chen L, Lewis NL, Nagarajan H, Sarkaria V, Kumar A, Wolozny D, Colao J, Jacobson E, Tian Y, O'Meally RN, Krag SS, Cole RN, Palsson BØ, Zhang H, Betenbaugh M. (2012) Proteomic analysis of Chinese hamster ovary cells. J Proteome Res. 11:5265-5276.

Becker J, Hackl M, Rupp O, Jakobi T, Schneider J, Szczepanowski R, Bekel T, Borth N, Goesmann A, Grillari J, Kaltschmidt C, Noll T, Puhler A, Tauch A, Brinkrolf K. (2011) Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. J Biotechnol. 156:227-235.

Bennett S. (2004) Solexa ltd. Pharmacogenomics. 5:433-438.

Berlec A, Strukelj B. (2013) Current state and recent advances in biopharmaceutical production in Escherichia coli, yeasts and mammalian cells. J Ind Microbiol Biot. 40:257-274.

BioIT World. (2014) Illumina announces the thousand dollar genome. Bio-IT World.

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

Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE, The Mouse Genome Database Group. (2014) The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. Nucleic Acids Res. 42:D810-817.

Boland JF, Chung CC, Roberson D, Mitchell J, Zhang X, Im KM, He J, Chanock SJ, Yeager M, Dean M. (2013) The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing. Hum Genet. 132:1153-1163.

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

Bosques CJ, Collins BE, Meador JW, Sarvaiya H, Murphy JL, DelloRusso G, Bulik DA, Hsu IH, Washburn N, Sipsey SF, Myette JR, Raman R, Shriver Z, Sasisekharan R, Venkataraman G. (2010) Chinese hamster ovary cells can produce galactose-α-1,3-galactose antigens on proteins. Nat Biotechnol. 28:1153-1156.

Brent, MR. (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet. 9:62-73.

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

Brooks SA, Dwek MV, Schumacher U. (2002) Functional and molecular glycobiology. BIOS Scientific Publishers Limited, Oxford, UK.

Butler M. (2005) Animal cell cultures: recent achievements and perspectives in the production of biopharmaceuticals. Appl Microbiol Biot. 68:283-291.

Cao Y, Kimura S, Itoi T, Honda K, Ohtake H, Omasa T. (2012) Construction of BAC-based physical map and analysis of chromosome rearrangement in Chinese hamster ovary cell lines. Biotechnol Bioeng. 109:1357-1367.

Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. (2012) Pacific biosciences sequencing technology for genotyping and variation and discovery in human data. BMC Genomics. 13:375-382.

Cervera L, Gutierrez-Granados S, Martinez M, Blanco J, Godia F, Segura MM. (2013) Generation of HIV-1 Gag VLPs by transient transfection of HEK 293 suspension cell culture using an optimized animal-derived component free medium. J Biotechnol. 166:152-165.

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J, Eichler EE. (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature. 517:608-611.

Chen N, Koumpouras GC, Polizzi KM, Kontoravdi C. (2012) Genome-based kinetic modeling of cytosolic glucose metabolism in industrially relevant cell lines: Saccharomyces cerevisiae and Chinese hamster ovary cells. Bioprocess Biosyst Eng. 35:1023-1033.

Chen P, Harcum SW. (2006) Effects of elevated ammonium on glycosylation gene expression in CHO cells. Metab Eng. 8:123-132.

Chong WPK, Reddy SG, Yusufi FNK, Lee DY, Wong NSC, Heng CK, Yap MGS, Ho YS. (2010) Metabolomics-driven approach for the improvement of Chinese hamster ovary cell growth: overexpression of malate dehydrogenase II. J Biotechnol. 147:116-121.

Chowdhury R, Chowdhury A, Maranas CD. (2015) Using gene essentiality and synthetic lethality information to correct yeast and CHO cell genome-scale models. Metabolites. 5:536-570.

Chung CH, Mirakhur B, Chan E, Le QT, Berlin J, Morse M, Murphy BA, Satinover SM, Hosen J, Mauro D, Slebos RJ, Zhou Q, Gold D, Hatley T, Hicklin DJ, Platts-Mills TAE. (2008) Cetuximab-induced anaphylaxis and IgE specific for galactose-alpha-1,3-galactose. N Engl J Med. 358:1109-1117.

Clarke C, Henry M, Doolan P, Kelly S, Aherne S, Sanchez N, Kelly P, Kinsella P, Breen L, Madden SF, Zhang L, Leonard M, Clynes M, Meleady P, Barron N. (2012) Integrated miRNA, mRNA and protein expression analysis reveals the role of post-transcriptional regulation in controlling CHO cell growth rate. BMC Genomics, 13:656.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 4:265-270.

Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BØ. (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature. 429:92-96.

Crea F, Sarti D, Falciani F, Al-Rubeai M. (2006) Over-expression of hTERT in CHOK1 results in decreased apoptosis and reduced serum dependency. J Biotechnol. 121:109-123.

Damerla RR, Chatterjee B, Li Y, Francis RJB, Fatakia SN, Lo CW. (2014) Ion Torrent sequencing for conducting genome-wide scans for mutation mapping analysis. Mamm Genome. 25:120-128.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. (2011) The variant call format and VCF tools. Bioinformatics. 27:2156-2158.

Datta P, Linhardt RJ, Sharfstein ST. (2013). An 'omics approach towards CHO cell engineering. Biotechnol Bioeng. 110:1255-1271.

Davies SL, Lovelady CS, Grainger RK, Racher AJ, Young RJ, James DC. (2013) Functional heterogeneity and heritability in CHO cell populations. Biotechnol Bioeng. 110:260-274.

Dell A, Morris HR, Greer F, Redfern JM, Rogers ME, Weisshaar G, Hiyama J, Renwick AGC. (1991) Fast-atom-bombardment mass spectrometry of sulphated oligosaccharides from ovine lutropin. Carbohyd Res. 209:33-50.

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

Dickson AJ. (2014) Enhancement of production of protein biopharmaceuticals by mammalian cell cultures: the metabolomics perspective. Curr Opin Biotechnol. 30:73-79.

Dietmair S, Nielsen LK, Timmins NE. (2012) Mammalian cells as biopharmaceutical production hosts in the age of omics. Biotechnol J. 7:75-89.

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. PNAS. 104:1777-1782.

Eid J, Fehr A, Gray J, Luong K, Lyle F, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, deWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, et al. (2009) Real-time DNA sequencing from single polymerase molecules. Science. 323:133-138.

Emrich CA, Tian HJ, Medintz IL, Mathies RA. (2002) Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. Anal Chem. 74:5076-5083.

English BP, Min W, van Oijen AM, Lee KT, Luo G, Sun H, Cherayil BJ, Kou SC, Xie XS. (2006) Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. Nat Chem Biol. 2:87-94.

Everett MV, Grau ED, Seeb JE. (2011) Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. Mol Ecol Resour. 11:93-108.

Farrell A, McLoughlin N, Milne JJ, Marison IW, Bones J. (2014) Application of multi-omics techniques for bioprocess design and optimization in Chinese hamster ovary cells. J Proteome Res. 13:3144-3159.

Fassler J, Cooper P. (2008) National Center for Biotechnology Information (US) (Ed.), BLAST® Help, National Center for Biotechnology Information (US), Bethesda. BLAST Glossary.

Feist AM, Palsson BØ. (2013) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol. 26:659-667.

Ferrari ML, Gomez GA, Maccioni HJF. (2012) Spatial organization and stoichiometry of n-terminal domain-mediated glycosyltransferase complexes in Golgi membranes determined by fret microscopy. Neurochem Res. 37:1325-1334.

Flicek P, Birney E. (2009) Sense from sequence reads: methods for alignment and assembly. Nat Methods. 6:S6-S12.

Flosberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 7:461-465.

Galan M, Guivier E, Caraux, Charbonnel N, Cosson JF. (2010) A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. BMC Genomics. 11:296.

Gatti MD, Wlaschin KF, Nissom PM, Yap M, Hu WS. (2007) Comparative transcriptional analysis of mouse hybridoma and recombinant Chinese hamster ovary cells undergoing butyrate treatment. J Biosci Bioeng. 103:82-91.

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

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera, Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Venter JC, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtel K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bouffard GG, NHGRI, de Jong PJ, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 428:493-521.

Glenn TC. (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour. 11:759-769.

Goh JSY, Liu Y, Chan KF, Wan C, Teo G, Zhang P, Zhang Y, Song Z. (2014) Producing recombinant therapeutic glycoproteins with enhanced sialylation using CHO-gmt4 glycosylation mutants. Bioengineered. 5:1–5.

Griffin TJ, Seth G, Xie H, Bandhakavi S, Hu WS. (2007) Advancing mammalian cell culture engineering using genome-scale technologies. Trends Biotechnol. 25:401-408.

Hackl M, Jadhav V, Jakobi T, Rupp O, Brinkrolf K, Goesmann A, Pühler A, Noll T, Borth N, Grillari J. (2012) Computational identification of microRNA gene loci and precursor microRNA sequences in CHO cell lines. J Biotechnol. 158:151-155.

Hackl M, Jakobi T, Blom J, Doppmeier D, Brinkrolf K, Szczepanowski R, Bernhart S, Siederdissen CH, Bort JAH, Wieser M, Kunert R, Jeffs S, Hofacker IL, Goesmann A, Pühler A, Borth N, Grillari J. (2011) Next-generation sequencing of the Chinese hamster ovary microRNA transcriptome: identification, annotation and profiling of microRNAs as targets for cellular engineering. J Biotechnol. 153:62-75.

Hammond S, Kaplarevic M, Borth N, Betenbaugh MJ, Lee KH. (2012) Chinese hamster genome database: an online resource for the CHO community at www.CHOgenome.org. Biotechnol Bioeng. 109:1353-1356.

Hammond S, Swanberg FC, Kaplarevic M, Lee KH. (2011) Genomic sequencing and analysis of a Chinese hamster ovary cell line using Illumina sequencing technology. BMC Genomics. 12:67.

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

Hassinen A, Kellokumpu S. (2014) Organizational interplay of Golgi N-glycosyltransferases involves organelle microenvironment-dependent transitions between enzyme homo- and heteromers. J Biol Chem. 289:26937-26948.

Hayduk EJ, Choe LH, Lee KH. (2004) A two-dimensional electrophoresis map of Chinese hamster ovary cell proteins based on fluorescence staining. Electrophoresis. 25:2545-2556.

Hayduk EJ, Lee KH. (2005) Cytochalasin D can improve heterologous protein productivity in adherent Chinese hamster ovary cells. Biotechnol Bioeng. 90:354-364.

Heffner KM, Hizal DB, Kumar A, Shiloach J, Zhu J, Bowen MA, Betenbaugh MJ. (2014) Exploiting the proteomics revolution in biotechnology: from disease and antibody targets to optimizing bioprocess development. Curr Opin Biotechnol. 30:80-86.

Higuchi M, Oheda M, Kuboniwa H, Tomonoh K, Shimonaka Y, Ochi N. (1992) Role of sugar chains in the expression of the biological-activity of human erythropoietin. J Biol Chem. 267:7703-7709.

Hoffman JI, Tucker R, Bridgett SF, Clark MS, Forcada J, Slate J. (2012) Rates of assay success and genotyping error when single nucleotide polymorphism genotyping in non-model organisms: a case study in the Antarctic fur seal. Mol Ecol Resour. 12:861-872.

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

Hossler P, McDermott S, Racicot C, Chumsae C, Raharimampionona H, Zhou Y, Ouellette D, Matuck J, Correia I, Fann J, Li J. (2014) Cell culture media supplementation of uncommonly used sugars sucrose and tagatose for the targeted shifting of protein glycosylation profiles of recombinant protein therapeutics. Biotechnol Prog. 30:1419-1431.

Huntzinger E, Izaurralde E. (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet. 12:99-110.

Ikehara Y, Sato T, Niwa T, Nakamura S, Gotoh M, Ikehara SK, Kiyohara K, Aoki C, Iwai T, Nakanishi H, Hirabayashi J, Tatematsu M, Narimatsu H. (2006) Apical Golgi localization of N,N'-diacetyllactosediamine synthase, beta4GalNAc-T3, is responsible for LacdiNAc expression on gastric mucosa. Glycobiology. 16:777-785.

Imai-Nishiya H, Mori K, Inoue M, Wakitani M, Iida S, Shitara K, Satoh M. (2007) Double knockdown of a1,6-fucosyltransferase (FUT8) and GDP-mannose 4,6-dehydratase (GMD) in antibody-producing cells: a new strategy for generating fully non-fucosylated therapeutic antibodies with enhanced ADCC. BMC Biotechnol. 7:84-96.

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

Jayapal KP, Wlaschin KF, Hu WS, Yap MGS. (2007) Recombinant protein therapeutics from CHO cells - 20 years and counting. Chem Eng Prog. 103:40-47.

Jenkins N, Meleady P, Tyther R, Murphy L. (2009) Strategies for analysing and improving the expression and quality of recombinant proteins made in mammalian cells. Biotechnol Appl Bioc. 53:73-83.

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

Johnson KC, Yongky A, Vishwanathan N, Jacob NM, Jayapal KP, Goudar CT, Karypis G, Hu WS. (2014) Exploring the transcriptome space of a recombinant BHK cell line through next generation sequencing. Biotechnol Bioeng. 111:770-781.

Kanda Y, Imai-Nishiya H, Kuni-Kamochi R, Mori K, Inoue M, Kitajima-Miyama K, Okazaki A, Iida S, Shitara K, Satoh M. (2007) Establishment of a GDP-mannose 4,6-dehydratase (GMD) knockout host cell line: A new strategy for generating completely non-fucosylated recombinant therapeutics. J Biotechnol. 130:300–310.

Kanehisa M, Goto S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27-30.

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

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

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 26:2204-2207.

Keohavong P, Thilly WG. (1989) Fidelity of DNA polymerases in DNA amplification. P Natl Acad Sci USA. 86:9253-9257.

Kildegaard HF, Baycin-Hizal D, Lewis NL, Betenbaugh MJ. (2013) The emerging CHO systems biology era: harnessing the ‘omics revolution for biotechnology. Curr Opin Biotechnol. 24:1102-1107.

Kim JY, Kim Y, Lee GM. (2012) CHO cells in biotechnology for production of recombinant proteins: current state and further potential. Appl Microbiol Biotechnol. 93:917-930.

Klanert G, Jadhav V, Chanoumidou K, Grillari J, Borth N, Hackl M. (2014) Endogenous microRNA clusters outperform chimeric sequence clusters in Chinese hamster ovary cells. Biotechnol J. 9:538-544.

Kochanowski N, Blanchard F, Cacan R, Chirat F, Guedon E, Marc A, Goergen JL. (2008) Influence of intracellular nucleotide and nucleotide sugar contents on recombinant interferon-γ glycosylation during batch and fed-batch cultures of CHO cells. Biotechnol Bioeng. 100:721-733.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Phillippy AM. (2012) Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat Biotechnol. 30:693-700.

Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW. (2010) Real-time DNA sequencing from single molecule polymerase molecules. Method Enzymol. 472:431-455.

Koster H, Tang K, Fu DJ *et al.* (1996) A strategy for rapid and efficient DNA sequencing by mass spectrometry. Nat Biotechnol. 14:1123-1128.

Koutny L, Schmalzing D, Salas-Solano O, El-Difrawy S, Adourian A, Buonocore S, Abbey K, McEwan P, Matsudaira P, Ehrlich D. (2000) Eight hundred base sequencing in a microfabricated electrophoretic device. Anal Chem. 72:3388-3391.

Kowalczyk SW, Wells DB, Aksimentiev A, Dekker C. (2012) Slowing down DNA translocation through a nanopore in lithium chloride. Nano Lett. 12:1038-1044.

Krambeck FJ, Bennum SV, Narang S, Choi S, Yarema KJ, Betenbaugh MJ. (2009) A mathematical model to derive N-glycan structures and cellular enzyme activities from mass spectrometric data. Glycobiology. 19:1163-1175.

Krambeck FJ, Betenbaugh MJ. (2005) A Mathematical Model of N-Linked Glycosylation. Biotechnol Bioeng. 92:711-728.

Kremkow B, Lee KH. (2013) Next-generation sequencing technologies and their potential impact on CHO cell-based biomanufacturing. Pharm Bioprocess. 1:455-465.

Kremkow BG, Baik JY, MacDonald ML, Lee KH. (2015) CHOgenome.org 2.0: Genome resources and website updates. Biotechnol J. 10:931-938.

Kremkow BG, Lee KH. (2015) Sequencing technologies for animal cell culture research. Biotechnol Lett. 37:55-65.

Kremkow BG, Lee KH. (Submitted) Glyco-Mapper: A Chinese hamster ovary (CHO) genome-specific glycosylation prediction tool.

Krol A. (2014) What you need to know about Illumina's new sequencers. Bio-IT World

La Merie Business Intelligence. (2013) Blockbuster biologics 2012. R&D Pipeline News. 7:2-28.

La Merie Business Intelligence. (2016) Blockbuster biologics 2015. R&D Pipeline News. 10:3-42.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond, C, Rosetti M, Santos R, Sheridan A, Sougnez C, Strange-Thomann N, Stojanovic N, Subramanian A, Wyman D, *et al.* (2001) Initial sequencing and analysis of the human genome. Nature. 409:860-921.

Lau KS, Partridge EA, Grigorian A, Silvescu CI, Reinhold VN, Demetriou M, Dennis JW. (2007) Complex N-glycan number and degree of branching cooperate to regulate cell proliferation and differentiation. Cell. 129:123-134.

Laulederkind SJ, Hayman GT, Wang SJ, Smith JR, Lowry TF, Nigam R, Petri V, de Pons J, Dwinell MR, Shimoyama M, Munzenmaier DH, Worthey EA, Jacob HJ. (2013) The Rat Genome Database 2013 – data, tools, and users. Brief Bioinform. 14:520-526.

Lemay MA, Henry P, Lamb CT, Robson KM, Russello MA. (2013) Novel genomic resources for a climate change sensitive mammal: characterization of the American pika transcriptome. BMC Genomics. 14:311.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078-2079.

Li RQ, Li YR, Kristiansen K, Wang J. (2008) SOAP: short oligonucleotide alignment program. Bioinformatics. 24:713-714.

Lis H, Sharon N. (1998) Lectins: carbohydrate-specific proteins that mediate cellular recognition. Chem Rev. 98:637-674.

Liu G, Neelamegham S. (2014) A Computational Framework for the Automated Construction of Glycosylation Reaction Networks. PLOS ONE. 9:e100939.

Liu L, Li YH, Li SL, Hu N, He Y, Pong R, Lin D, Lu L, Law M. (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012:1-11.

Luo J, Vijayasankaran N, Autsen J, Santuray R, Hudson T, Amanullah A, Li F. (2012) Comparative metabolite analysis to understand lactate metabolism shift in Chinese hamster ovary cell culture process. Biotechnol Bioeng. 109:146-156.

Maccani A, Hackl M, Leitner C, Steinfellner W, Graf AB, Tatto NE, Karbiener M, Scheideler M, Grillari J, Mattanovich D, Kunert R, Borth N, Grabherr R, Ernst W. (2014) Identification of microRNAs specific for high producer CHO cell lines using steady-state cultivation. Appl Microbiol Biotechnol. 98:7535-7548.

Macher BA, Galili U. (2008) The Gal alpha 1,3Gal beta 1,4GlcNAc-R(alpha-Gal) epitope: a carbohydrate of unique evolution and clinical relevance. Biochim Biophys Acta. 1780:75-88.

Malhotra R, Wormald MR, Rudd PM, Fischer PB, Dwer RA, Sim RB. (1995) Glycosylation changes of IgG associated with rheumatoid arthritis can activate complement via the mannose-binding protein. Nat Med. 1:237-243.

Malphettes L, Freyvert Y, Chang J, Liu PQ, Chan E, Miller JC, Zhou Z, Nguyen T, Tsai C, Snowden AW, Collingwood TN, Gregory PD, Cost GJ. (2010) Highly Efficient Deletion of FUT8 in CHO Cell Lines Using Zinc-Finger Nucleases Yields Cells That Produce Completely Nonfucosylated Antibodies. Biotechnol Bioeng. 130:300–310.

Mardis ER. (2008) Next-generation DNA sequencing methods. Annu Rev Genom Hum G. 9:387-402.

Mariño K, Bones J, Kattla JJ, Rudd PM. (2010) A systematic approach to protein glycosylation analysis: a path through the maze. Nat Chem Biol. 6:713-723.

Maszczak-Seneczko D, Sosicka P, Olczak T, Jakimowicz P, Majkowski M, Olczak M. (2013) UDP-N-acetylglucosamine Transporter (SLC35A3) Regulates Biosynthesis of Highly Branched N-glycans and Keratan Sulfate. J Biol Chem. 288:21850–21860.

McCarthy A. (2010) Third generation DNA sequencing: Pacific Biosciences' Single Molecule Real Time technology. Chem Biol. 17:675-676.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297-1303.

McQuilton P, St Pierre SE, Thurmond J, FlyBase Consortium. (2012) FlyBase 101--the basics of navigating FlyBase. Nucleic Acids Res. 40:D706-714.

Mechref Y, Muzikar J, Novotny MV. (2005) Comprehensive assessment of N-glycans derived from a murine monoclonal antibody: A case for multimethodological approach. Electrophoresis. 26:2034-2046.

Meleady P, Hoffrogge R, Henry M, Rupp O, Bort JH, Clarke C, Brinkrolf K, Kelly S, Müller B, Doolan P, Hackl M, Beckmann TF, Noll T, Grillari J, Barron N, Pühler A, Clynes M, Borth N. (2012) Utilization and evaluation of CHO-specific sequence databases for mass spectrometry based proteomics. Biotechnol Bioeng. 109:1386-1394.

Melmer M, Strangler T, Premstaller A, Lindner W. (2011) Comparison of hydrophilic-interaction, reversed-phase and porous graphitic carbon chromatography for glycan analysis. J Chromatogr A. 1218:118-123.

Merriman B, Rothberg JM, Ion Torrent R&D Team. (2012) Progress in Ion Torrent semiconductor chip based sequencing. Electrophoresis. 33:3397-3417.

Metzker ML. (2010) Sequencing technologies – the next generation. Nat Rev. 11: 31-46.

Miller JR, Koren S, Sutton G. (2010) Assembly algorithms for next-generation sequencing data. Genomics. 95:315-327.

Naik AD, Menegatti S, Gurgel PV, Carbonell RG. (2011) Performance of hexamer peptide ligands for affinity purification immunoglobulin G from commercial cell culture media. J Chromatogr A. 1218:1691-1700.

Nam JH, Zhang F, Ermonval M, Linhardt RJ, Sharfstein ST. (2008) The effects of culture conditions on the glycosylation of secreted human placental alkaline phosphatase produced in Chinese hamster ovary cells. Biotechnol Bioeng. 100:1178-1192.

Naso MF, Tam SH, Scallon BJ, Raju TS. (2010) Engineering host cell lines to reduce terminal sialylation of secreted antibodies. mAbs. 2:519-527.

National Center for Biotechnology Information (US). (2013) The NCBI Handbook. National Center for Biotechnology Information (US), Bethesda.

Nissom PM, Sanny A, Kok YJ, Hiang YT, Chuah SH, Shing TK, Lee YY, Wong KTK, Hu WS, Yap MGS, Philp R. (2006) Transcriptome and proteome profiling to understanding the biology of high productivity CHO cells. Mol Biotechnol. 34:125-140.

North SJ, Huang HH, Sundaram S, Jang-Lee J, Etienne T, Trollope A, Chalabi S, Dell A, Stanley P, Haslam SM. (2010) Glycomics profiling of Chinese hamster ovary cell glycosylation mutants reveals N-glycans of a novel size and complexity. J Biol Chem. 285:5759-5775.

Nyberg GB, Balcarcel RR, Follstad BD, Stephanopoulos G, Wang DIC. (1998) Metabolic effects on recombinant interferon-y glycosylation in continuous culture of Chinese hamster ovary cells. Biotechnol Bioeng. 62:336-347.

O'Neill RA. (1996) Enzymatic release of oligosaccharides from glycoproteins for chromatographic and electrophoretic analysis. J Chromatogr A. 720:201-215.

Omasa T, Cao YH, Park JY, Takagi Y, Kimura S, Yano h, Honda K, Asakawa S, Shimizu N, Ohtake H. (2009) Bacterial artificial chromosome library for genome-wide analysis of Chinese hamster ovary cells. Biotechnol Bioeng. 104:986-994.

Onitsuka M, Kim WD, Ozaki H, Kawaguchi A, Honda K, Kajiura H, Fujiyama K, Asano R, Kumagai I, Ohtake H, Omasa T. (2012) Enhancement of sialylation on humanized IgG-like bispecific antibody by overexpression of α-2,6-sialyltransferase derived from Chinese hamster ovary cells. Biotechnol Prod Proc Eng. 94:69–80.

Orlova NA, Orlov AV, Vorobiev II. (2012) A modular assembly cloning technique (aided by the BIOF software tool) for seamless and error-free assembly of long DNA fragments. BMC Res Notes. 5:303.

Ouyang A, Bennett P, Zhang A, Yang ST. (2007) Affinity chromatographic separation of secreted alkaline phosphatase and glucoamylase using reactive dyes. Process Biochem. 42:561-569.

Partridge MA, Davidson MM, Hei TK. (2007) The complete nucleotide sequence of Chinese hamster (Cricetulus griseus) mitochondrial DNA. DNA Seq. 18:341-346.

Pascoe DE, Arnott D, Papoutsakis ET, Miller WM, Andersen DC. (2007) Proteome analysis of anti body-producing CHO cell lines with different metabolic profiles. Biotechnol Bioeng. 98:391-410.

Pharmaceutical Research and Manufacturers of America. (2015) 2015 biopharmaceutical research industry profile. PhRMA. 1-76.

Picotti P, Aebersold R. (2012) Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods. 9:555-566.

Prater BD, Connelly HM, Qin Q, Cockrill SL. (2009) High-throughput immunoglobulin G N-glycan characterization using rapid resolution reverse-phase chromatography tandem mass spectrometry. Anal Biochem. 385:69-79.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 13:341.

Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, Marth GT, Stein L, Flicek P, Yandell M, Eilbeck K. (2010) A standard variation file format for human genome sequences. Genome Biol. 11:R88.

Reinders J, Pasazkowski J. (2010) Bisulfite methylation profiling of large genomes. Epigenomics. 2:209-220.

Ressom H, Natarajan P, Varghese RS, Musavi MT. (2005) Applications of fuzzy logic in genomics. Fuzzy Set Syst. 152:125-138.

Rhee M, Burns MA. (2006) Nanopore sequencing technology: research trends and applications. Trends Biotechnol. 24:580-586.

Rivinoja A, Hassinen A, Kokkonen N, Kauppila A, Kellokumpu S. (2009) Elevated Golgi pH impairs terminal N-glycosylation by inducing mislocalization of Golgi glycosyltransferases. J Cell Physiol. 220:144-154.

Roberts RJ, Carneiro MO, Schatz MC. (2013) The advantages of SMRT sequencing. Genome Biol. 14:405.

Ronda C, Pedersen LE, Hansen HG, Kallehauge TB, Betenbaugh MJ, Nielsen AT, Kildegaard HF. (2014) Accelerating genome editing in CHO cells using CRISPR Cas9 and CRISPy, a web-based target finding tool. Biotechnol Bioeng. 111:1604-1616.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. (2013) Characterizing and measuring bias in sequence data. Genome Biol. 14:51-70.

Rothberg JM, Leamon JH. (2008) The development and impact of 454 sequencing. Nat Biotechnol. 26:1117-1124.

Sandra K, Vandenheede I, Sandra P. (2014) Modern chromatographic and mass spectrometric techniques for protein biopharmaceutical characterization. J Chromatogr A. 1335:81-103.

Sanger F, Nicklen S, Coulson AR. (1977) DNA sequencing with chain-terminating inhibitors. P Natl Acad Sci USA. 74:5463-5467.

Schadt EE, Turner S, Kasarskis A. (2010) A window into third-generation sequencing. Hum Mol Genet. 19:R227-R240.

Scherer WF, Syverton JT, Gey GO. (1953) Studies on the propagation in vitro of poliomyelitis viruses 4 – viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix. J Exp Med. 97:695-710.

Schuster SC. (2008) Next-generation sequencing transforms today's biology. Nat Methods. 5:16-18.

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

Selvarasu S, Ho YS, Chong WPK, Wong NSC, Yusufi FNK, Lee YY, Yap MGS, Lee DY. (2012) Combined *in silico* modeling and metabolomics analysis to characterize fed-batch CHO cell culture. Biotechnol Bioeng. 109:1415-1429.

Selvarasu S, Karimi IA, Ghim GH, Lee DY. (2010) Genome-scale modeling and *in silico* analysis of mouse cell metabolic network. Mol Biosyst. 6:152-161.

Service RF. (2006) Gene sequencing - The race for the $1000 genome. Science. 311: 1544-1546.

Sharon D, Tilgner H, Grubert F, Snyder M. (2013) A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 31:1009-1014.

Shaw TI, Srivastava A, Chou WC, Liu L, Hawkinson A, Glenn TC, Adams R, Schountz T. (2012) Transcriptome sequencing and annotation for the Jamaican fruit bat (*Artibeus jamaicensis*). Plos One. 7:1-12.

Sheikh K, Forster J, Nielsen LK. (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus*. Biotechnol Progr. 21:112-121.

Shendure J, Ji H. (2008) Next-generation DNA sequencing. Nat Biotechnol. 26:1135-1145.

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

Shlomi T, Cabili MN, Herrgård MJ, Palsson BØ, Ruppin E. (2008) Network-based prediction of human tissue-specific metabolism. Nat Biotechnol. 26:1003-1010.

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

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. (2009) JBrowse: a next-generation genome browser. Genome Res. 19:1630-1638.

Sokhansanj BA, Datta S, Hu X. (2009) Scalable dynamic fuzzy biomolecular network models for large scale biology. Fuzzy Systems in Bio. 1:235-255.

Spahn PN, Hansen AH, Hansen HG, Arnsdorf J, Kildegaard HF, Lewis NE. (2016) A Markov chain model for N-linked protein glycosylation – towards a low parameter tool for model-driven glycoengineering. Metabol Eng. 33:52-66.

St. Amand MM, Radhakrishnan D, Robinson AS, Ogunnaike BA. (2014) Identification of manipulated variables for a glycosylation control strategy. Biotechnol Bioeng. 111:1957-1970.

Suresh BV, Roy R, Sahu K, Misra G, Chattopadhyay D. (2014) Tomato genomic resources database: an integrated repository of useful tomato genomic information for basic and applied research. PLoS One. 9:e86387.

Taniguchi N, Honke K, Fukuda M. (2002) Handbook of glycosyltransferases and related genes. Springer, New York, New York, USA.

Tateno H, Uchiyama N, Kuno A, Togayachi A, Sato T, Narimatsu H, Hirabayashi J. (2007) A novel strategy for mammalian cell surface glycome profiling using lectin microarray. Glycobiology. 17:1138-1146.

Thiele I, Palsson BØ. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 5:93-121.

Thiele I, Price ND, Vo TD, Palsson BØ. (2005) Candidate metabolic network states in human mitochondria: impact of diabetes, ischemia and diet. J Biol Chem. 280:11683-11695.

Thomson T, Lin HF. (2009) The Biogenesis and function of PIWI proteins and piRNAs: progress and prospect. Annu Rev Cell and Dev Bi. 25:355-376.

Trapnell C, Salzberg SL. (2009) How to map billions of short reads onto genomes. Nat Biotechnol. 27:455-457.

Trummer E, Fauland K, Seidinger S, Schriebl, Lattenmayer C, Kunert R, Vorauer-Uhl K, Weik R, Borth N, Katinger H, Muller D. (2006) Process parameter shifting part I. Effect of DOT, pH and temperature on the performance of Epo-Fc expressing CHO cells cultivated in controlled batch bioreactors. Biotechnol Bioeng. 94:1033-1044.

Tsukahara M, Aoki A, Kozono K, Maseki Y, Fukuda Y, Yoshida H, Kobayashi K, Kakitani M, Tomizuka K, Tsumura H. (2006) Targeted disruption of α-1,6-fucosyltransferase (FUT8) gene by homologous recombination in Chinese hamster ovary (CHO) cells. Animal Cell Technol. 14:175-183.

Umaña P, Bailey JE. (1997) A Mathematical Model of N-linked Glycoform Biosynthesis. Biotechnol Bioeng. 55:890-908.

US DHHS FDA. (2015) Quality considerations in demonstrating biosimilarity of a therapeutic protein product to a reference product. 1:1-19.

Valente KN, Schaefer AK, Kempton HR, Lenhoff AM, Lee KH. (2014) Recovery of Chinese hamster ovary host cell proteins for proteomic analysis. Biotechnol J. 9:87-99.

Van Dijk ADJ, Bosch D, ter Braak CJF, van der Krol AR, van Ham RCHJ. (2008) Predicting sub-Golgi localization of type II membrane proteins. Bioinformatics. 24:1779-1786.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson D, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Miklos GLG, Nelson C, Broder S, *et al.* (2001) The sequence of the human genome. Science. 291:1304-1351.

Vermassen T, Van Praet C, Vanderschaeghe D, Maenhout T, Lumen N, Callewaert N, Hoebeke P, Van Belle S, Rottey S, Delanghe J. (2014) Capillary electrophoresis of urinary prostate glycoproteins assists in the diagnosis of prostate cancer. Electrophoresis. 35:1017-1024.

Vishwanathan N, Le H, Le T, Hu WS. (2014) Advancing biopharmaceutical process science through transcriptome analysis. Curr Opin Biotechnol. 30:113-119.

Wall DP, Fraser HB, Hirsh AE. (2003) Detecting putative orthologs. Bioinformatics. 19:1710-1711.

Walsh G, Jefferis R. (2006) Post-translational modifications in the context of therapeutic proteins. Nat Biotechnol. 24:1241-1252.

Walsh G. (2007) Pharmaceutical biotechnology. John Wiley & Sons Inc. 1:1-11.

Walsh G. (2010) Biopharmaceutical benchmarks 2010. Nat Biotechnol. 28:917-924.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520-562.

Weikert S, Papac D, Briggs J, Cowfer D, Tom S, Gawlitzek M, Lofgren J, Mehta S, Chisholm V, Modi N, Eppler S, Carroll K, Chamow S, Peers D, Berman P, Krummen L. (1999) Engineering Chinese hamster ovary cells to maximize sialic acid content of recombinant glycoproteins. Nat Biotechnol. 17:1116-1121.

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

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

Wong NSC, Wati L, Nissom PM, Feng HT, Lee MM, Yap MGS. (2010) An investigation of intracellular glycosylation activities in CHO cells: Effects of nucleotide sugar precursor feeding. Biotechnol Bioeng. 107:321-326.

Wuest DM, Harcum SW, Lee KH. (2012) Genomics in mammalian cell culture bioprocessing. Biotechnol Adv. 30:629-638.

Wurm FM, Hacker D. (2011) First CHO genome. Nat Biotechnol. 29:718-720.

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

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

Yandell M, Ence D. (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 13:329-342.

Yang Z, Wang S, Halim A, Schulz MA, Frodin F, Rahman SH, Vester-Christensen MB, Behrens C, Kristensen C, Vakhrushev SY, Bennett EP, Wandall HH, Clausen H. (2015) Engineered CHO cells for production of diverse, homogeneous glycoproteins. Nat Biotechnol. 33:842-844.

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

Yee JC, Wlaschin KF, Chuah SH, Nissom PM, Hu WS. (2008) Quality assessment of cross-species hybridization of CHO transcriptome on a mouse DNA oligo microarray. Biotechnol Bioeng. 101:1359-1365.

Yin B, Gao Y, Chung CY, Yang S, Blake E, Stuczynski MC, Tang J, Kildegaard HF, Anderson MR, Zhang H, Betenbaugh MJ. (2015) Glycoengineering of Chinese hamster ovary cells for enhanced erythropoietin N-glycan branching and sialylation. Biotechnol Bioeng. 112:2343-2351.

Yu M, Hu Z, Pacis E, Vijayasankaran N, Shen A, Li F. (2011) Understanding the intracellular effect of enhanced nutrient feeding toward high titer antibody production process. Biotechnol Bioeng. 108:1078-1088.

Zamore PD, Haley B. (2005) Ribo-gnome: the big world of small RNAs. Science. 309:1519-1524.

Zanetta JP, Pons A, Richet C, Huet G, Timmerman P, Leroy Y, Bohin A, Bohin JP, Trinel PA, Poulain D, Hofsteenge J. (2004) Quantitative gas chromatography/mass spectrometry determination of C-mannosylation of tryptophan residues in glycoproteins. Anal Biochem. 329:199-206.

Zhang S, Wang RS, Zhang XS, Chen L. (2009) Fuzzy system methods in modeling gene expression and analyzing protein networks. Fuzzy Systems in Bio. 1:235-255.

Zhou X, Hong GY, Huang BB, Duan YM, Shen JY, Ni MW, Cong WT, Jin LT. (2014) Improved conditions for periodate/Schiff's base-based fluorescent staining of glycoproteins with dansylhydrazine in SDS-PAGE. Electrophoresis. 35:1439-1447.

Zhu J. (2012) Mammalian cell protein expression for biopharmaceutical production. Biotechnol Adv. 30:1158-1170.

Appendix A

COMPARISON OF 2ND AND 3RD GENERATION SEQUENCING TECHNOLOGIES AGAINST SANGER SEQUENCING

A.1 Preface

This appendix is adapted from Kremkow and Lee (2015) with permission (see Appendix D). This appendix presents a technical overview of traditional Sanger sequencing; the second generation sequencing technologies 454, Illumina, SOLiD, and Ion Torrent; and the third generation sequencing technologies PacBio and Nanopore Sequencing. The characteristics of each technology are summarized and compared. The potential impact these technologies may have on industrial biomanufacturing applications are considered.

A.2 Abstract

Over the last ten years, 2nd and 3rd generation sequencing technologies have made the use of genomic sequencing within the animal cell culture community increasingly commonplace. Each technology's defining characteristics are unique, including the sequencing cost, time, daily throughput, sequence read length, and occurrence of sequence errors. Given each sequencing technology's intrinsic advantages and disadvantages, the optimal technology for a given experiment depends on the particular experiment's objective. This appendix discusses the current characteristics of six next-generation sequencing technologies, compares the differences between them, and characterizes their relevance to the animal cell culture community. These technologies are continually improving, as evidenced by the recent achievement of the field's benchmark goal: sequencing a human genome for less than $1,000.

A.3 Introduction

Animal cell culture has progressed from the initial cultivation of the HeLa cell in the 1950's (Scherer et al. 1953) to the production of biopharmaceuticals in mammalian cells accounting for $86 billion in global sales in 2012 (Kremkow and Lee 2013; La Merie 2013). The number of animal cell culture-derived products continues to increase, progress that is accelerated by the commercial production of specialized cellular media, culturing equipment, and quantitative assay reagents (Naik et al. 2011, Cervera et al. 2013, Yu et al. 2011). Analytical technology is continually increasing in

capability and speed, while the size of the instrumentation is decreasing. These achievements are critical for future scientific advancements in visual microscopy (Baker 2010), genetic sequencing (Schuster 2008), and quantitative analysis (Picotti and Aebersold 2012). The last decade resulted in a marked increase in the number of DNA sequencing technologies and subsequent applications, enabling genetic sequencing experiments previously deemed impractical or impossible.

A.4 Sequencing Technologies

A.4.1 Sanger Sequencing Technology

The four nucleotides adenine (A), guanine (G), cytosine (C), and thymine (T) code for the cellular machinery required for growth and life. Since the 1970's, the Sanger sequencing chemistry (Sanger et al. 1977) has enabled researchers to successfully sequence these nucleotides. This method has since been improved upon but chain termination chemistry is still used to produce long read lengths ranging from 600 to 900 base pairs with a nucleotide substitution error rate between 0.01% and 0.1%. Despite the long, high-quality reads, Sanger sequencing is low-throughput, as significant investments of time and money are required to produce enough reads to sequence even a small genome; for example, to sequence one million bases would require approximately $2,000 and a minimum machine run time of 32 hours (Glenn 2011). For studies requiring a large number of sequenced bases, such as fully sequencing eukaryotic genomes, transcriptomes, and exomes, Sanger sequencing is impractical (Schuster 2008). The sequencing of biopharmaceutically-relevant cell line

genomes becomes increasingly impractical when the complexity and number of unique cell lines responsible for biopharmaceutical production is considered. Improved sequencing technologies with increased throughput and decreased cost were required to address these challenges.

A.4.2 Next-Generation Sequencing Technologies

Over the last ten years multiple new sequencing technologies have been developed. Each technology has distinctive sequencing characteristics, including sequencing errors, costs, time requirements, and sequence read length. A shared trait for one group of these technologies is sequence amplification, the production of multiple identical copies of a DNA sequence, and this group of technologies will be referred to as 2nd generation sequencing (SGS) technologies for the remainder of this appendix. Another group of sequencing technologies that does not require sequence amplification has been developed more recently and will be referred to as 3rd generation sequencing (TGS) technologies. Previous reviews of sequencing technologies have been published (Shendure and Ji 2008, Ansorge 2009, Metzker 2010, Schadt et al. 2010, Glenn 2011, Liu et al. 2012) and our goal in this appendix is to provide an update with an expanded, comprehensive comparison of these technologies' characteristics, specifically as applied to the animal cell technology community when possible. As there are a limited number of examples focused on biomanufacturing and technological bioprocesses, additional cases from the animal cell culture community will be used when necessary to further demonstrate these techniques.

Each sequencing technology provides distinct advantages and disadvantages originating from each technology's characteristics, with the most important sequencing technology characteristics being the (daily) experimental throughput, sequence assembly, cost, and experimental bias (Liu et al. 2012). The daily throughput, the number of nucleotides sequenced per day, is determined by the read length, number of reads generated per experiment, and run time per experiment. As the read length or number of reads generated increases, or the run time per experiment decreases, the experimental throughput increases. Read assembly, the compilation of the sequencing reads into the complete genome sequence, is dependent on the read length and the number of reads generated per experiment. The correlation between these variables is such that as the read length becomes shorter or the number of reads generated decreases in number, the correct assembly of the reads becomes increasingly difficult. As the quantity of sequenced nucleotides increases, a longer instrument run time is required, more reagents are needed, and the overall cost increases. The experimental sequence bias comprises coverage and error biases (Ross et al. 2013). Coverage bias originates from the uneven distribution of reads across the genome, possibly causing sections of the genome with decreased coverage to be inaccurate, while error bias originates from erroneous nucleotide substitutions, insertions, or deletions. As reads become shorter and less accurate, there is an increased probability of mismapping or misaligning the reads, particularly as the integrity of the nucleotide sequences decreases (Ross et al. 2013), further complicating the assembly process (McKenna et al. 2010). An optimal sequencing technology will have a large experimental throughput with no experimental bias and an accurate read assembly process for a low cost, yet no technology has all of these ideal

characteristics. Each sequencing technology will be described in the context of these primary characteristics as well as their advantages, disadvantages, and potential applications.

A.4.2.1 Second-Generation Sequencing (SGS) Technologies

SGS methods were developed to address the limited throughput of Sanger sequencing. The characteristics of four commonly used SGS technologies described in this section are compiled in Table A.1. Three of the sequencing technologies use fluorescence detection, while the fourth technology uses electrochemical detection. All of these technologies require amplification, achieve a much greater daily throughput than Sanger sequencing, and produce a large number of reads, albeit with shorter read lengths.

Table A.1: SGS Technology Characteristics

Common Name of SGS Tech-nology	Sequencer	Read Length (bases per read)	Cost[j] ($ per Mb)	Daily Through-put (Gb per day)	Run Time (days per run)	Reads (million per run)
454[a]	454 GS FLX+	450-900[g]	10	0.7	0.95	1[o]
Illumina[b]	Illumina HiSeq 2500[e]	100-150[h]	0.10	360-500[k]	5-6[m]	2,000-4,000[p]
SOLiD[c]	AB SOLiD 5500	35-75[i]	0.10	7-9	7[n]	0.7[q]
Ion Torrent[d]	Ion Torrent Ion Proton[f]	150-200	5	1-10[l]	2-4	60-80

[a]Website with sequencing data: http://454.com/products/gs-flx-system
[b]Website with sequencing data: www.illumina.com/systems/hiseq_2000_1000.ilmn
[c]Website with sequencing data: http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/solid-sequencing-chemistry.html
[d]Website with sequencing data: http://ioncommunity.lifetechnologies.com/community/intro
[e]Single/Dual flow cell
[f]Chip I – exome sequencing
[g]700 base average and 1,000 base maximum; 85% bases from reads > 500 bases, 45% bases from reads > 700 bases
[h]1x36 and 2x50 options available
[i]75x35 paired ends, 2x60 MP
[j]Adapted from (Liu et al. 2012) for 454, Illumina, and SOLiD
[k]Expected range for single flow cell, dual flow cell range is 720-950 Gb, whole range is 64-950 Gb
[l]Per run on Chip I (exome)
[m]Full run time for a single flow cell, but the complete range for all possible modes is 0.3-11 days
[n]1 lane for 1 genome at MP: 60 bp x 60 bp
[o]Using shotgun methods
[p]Single flow cell value, dual flow cell range: 4-8 billion, total range: 1.5-8 billion
[q]Per panel

A.4.2.1.1 454

Roche's 454 sequencing technology was the first introduced SGS technology and uses pyrosequencing (Rothberg and Leamon 2008). DNA is fragmented using nebulization and the fragments are amplified by emulsion PCR and attached to individual beads. Unlabeled nucleotides are added to the solution encapsulating the beads, enabling the production of up to one million DNA fragment copies per bead. Each bead is then deposited in one of the numerous wells located on the flat picotiter plate (PTP), to which a solution of substrates, enzymes (including DNA polymerase and luciferase), and additional unlabeled nucleotides is added. The four unlabeled nucleotides are flowed separately across the PTP and when a nucleotide is incorporated, a pyrophosphate molecule is released, initiating an enzymatic chain reaction and resulting in fluorescence. The fluorescence intensity within the PTP well is proportional to the number of nucleotides incorporated. Following each reaction, the PTP is washed and the sequencing cycle continues with another unlabeled nucleotide.

The average read length produced by the 454 technology ranges from 450 to 900 base pairs, with up to one million reads generated per run. The 454 technology is incapable of interpreting long stretches of the same nucleotide due to inconclusive assay measurements of high fluorescence intensity. These homopolymer nucleotide occurrences can cause insertion or deletion errors, which compose the majority of 454's approximate 1% error rate. Nucleotides following an insertion or deletion error are also affected because their location will be incorrectly shifted. For example, the 1% error rate equates to two incorrect nucleotides within the first 200 base pairs in a read. As each read is commonly longer than 450 base pairs, the positions of more than

50% of the nucleotides that compose each read are likely altered. A complete run can be finished in slightly less than 24 hours, resulting in a daily throughput of approximately 700 megabases (Mb) and an estimated cost per Mb of $10 (Glenn 2011). Compared to Sanger sequencing, 454 yields a greatly increased number of reads at a lower cost and a slightly reduced average read length for large genomic sequences; however, the increased occurrence of insertion and deletion errors reduces the accuracy of these reads. Within the animal cell technology community, 454 has been used to sequence portions of the Chinese hamster ovary (CHO) genome (Kantardjieff et al. 2009) and sequence transcripts from the CHO-K1 cell line (Becker et al. 2011); in another mammalian sequencing experiment, 454 has recently been used to identify single nucleotide polymorphisms (SNPs) in *Ochotona princeps* (Lemay et al. 2013). Potential future applications include additional transcriptomic sequencing experiments, genotyping polymorphic genes, and investigating desirable phenotypes to generate fundamental information that may further improve cellular performance within the animal cell technology community (Galan et al. 2010).

A.4.2.1.2 Illumina

Illumina's sequencing by synthesis (SBS) technology platform (Bennett 2004) is one of the most common sequencing platforms in use today. SBS preparation consists of random sequence fragmentation, ligation to an Illumina-specific adapter library, and multiple rounds of 'bridge' amplification, resulting in dense clusters of identical DNA fragments randomly located throughout a flow cell. The sequencing cycle is initiated by the simultaneous addition of primers, DNA polymerase enzymes, and the four fluorescently-labeled nucleotides. The DNA polymerase extends each

DNA fragment strand by one nucleotide and these terminal, fluorescently-labeled nucleotides emit a specific fluorescence, identifying the incorporated nucleotide. The fluorophore is removed by a chemical modification, enabling the continuation of the sequencing cycle.

Illumina's read lengths typically range between 100 and 150 base pairs, which is much shorter than the 454 read length. However, with an optimized preparation method and within five days, up to four billion reads per run can be generated. This can yield a maximum daily throughput of 500 gigabases (Gb), which is almost three orders of magnitude greater than 454's daily throughput. Due to the large number of short reads, complex and computationally efficient assembly algorithms are required for read assembly. The short read length is offset by the reduced cost per base (less than $0.10 per Mb), which is two orders of magnitude less than 454. Illumina's error rate is low, typically ranging between 0.1% and 0.5% (Ross et al. 2013) and primarily consists of substitution errors.

Current Illumina sequencing applications include genome sequencing, transcriptomic analysis, and small RNA discovery. The large number of reads generated from Illumina experiments enables the differentiation of a SNP from a substitution error. However, when genomic sequences that are longer than the read length are identical or highly similar originate from different genomic sequences, the reads may be incorrectly identified as copies, reads with SNPs, or substitution errors, rather than highly similar reads originating from unique genomic sequences. This misidentification is one problem Illumina's short read length causes for organisms with repetitious genome sequences. The repetitious sequences make *de novo* assembly of these genomes difficult without a reference genome and only slightly easier when a

reference genome is available. A large variety of research pursuits is achievable and within the animal cell technology community, Illumina has been used to sequence the Chinese hamster (Brinkrolf et al. 2013, Lewis et al. 2013) and CHO (Hammond et al. 2011, Xu et al. 2011) genomes, as well as the baby hamster kidney and CHO cell line transcriptomes (Johnson et al. 2014, Hackl et al. 2011). Overall, the Illumina platform is able to produce large amounts of high-quality data via a low-cost, high-throughput methodology.

A.4.2.1.3 SOLiD

The Sequencing by Oligonucleotide Ligation and Detection (SOLiD) sequencing technology platform uses magnetic bead-emulsion PCR for amplification, which is similar to 454's PCR technique previously described; however, the beads are deposited onto a flow cell instead of a PTP. SOLiD uses the enzyme DNA ligase and five unique, universal sequencing primers, where the primer attaches to the matching adapter and is extended by the attachment of a fluorescent 8-mer corresponding to the template fragment's first two nucleotides. The fluorophore is cleaved after the fluorescence is measured and the DNA ligase incorporates the 8-mer corresponding to the template fragment's next two nucleotides. This cycle of ligate, image, and cleave is repeated until the primer sequence is fully extended. The entire primer strand is then removed, the next primer is attached, and the ligate, image, and cleave cycle is repeated.

Unlike the other fluorescence-based methods, each fluorescent signal is representative of not one, but two nucleotides through a process called two-base coding. The fluorescent signal identifies four out of the sixteen potential nucleotide

pairs, but each nucleotide can only be correctly determined if the fluorescent signal from both primers associated with the specific nucleotide are measured. The combination of both fluorescent signals identifies the single, correct nucleotide pair combination. This process greatly reduces insertion and deletion errors, resulting in an error rate less than 0.1% that is primarily attributed to A-T bias. Read lengths for SOLiD range between 35 and 75 base pairs and each run can potentially yield between seven and nine Gb over the course of a full experiment, which requires up to seven days. This low error rate and short read length, the shortest reads in the SGS technology group, also yield many error-free reads. The cost per million bases is similar to that of Illumina and slightly more than $0.10 per Mb. Upon comparison of SOLiD's reads to a high-quality reference, the low error rate permits the differentiation between sequencing errors and SNPs with a high degree of certainty (Everett et al. 2011). Within the animal cell technology community, the current absence of high-quality reference sequence data for comparison increases the difficulty of SNP detection. Genome sequences with repeat regions greater than 75 bases in length are indistinguishable due to the short read lengths, one disadvantage shared by the SOLiD and Illumina platforms. SOLiD is not the SGS technology commonly used for animal cell culture sequencing due to the shorter read length and reduced throughput relative to the Illumina technology. However, this technology may be more commonly used in the future as more high-quality reference genomes are available for relevant animal cells.

A.4.2.1.4 Ion Torrent

Life Technologies' Ion Torrent requires sequence amplification, but unlike the other SGS technologies, Ion Torrent uses electrochemical detection, not fluorescence detection. Prior to sequencing, the DNA sample is fragmented and amplified using a method similar to bead-emulsion PCR. The beads are individually deposited into one of the 660 million wells on a chip (Merriman et al. 2012) and the chip is then flooded with a nucleotide solution. Upon the incorporation of a nucleotide, a hydrogen ion (H+) is released, thereby changing the pH of the well solution. The ion-sensitive bottom of the well measures this pH change and converts it to a voltage, which the chip measures quantitatively. The measured voltage correlates with the number of incorporated nucleotides. Specifically, the pH change and voltage proportionally increase in accordance with the number of integrated nucleotides. This process is repeated every fifteen seconds until sequencing is completed.

The maximum number of reads generated per Ion Torrent experiment is 70 million reads, with a short run time that ranges between two and four hours. The resulting daily throughput of approximately 10 Gb has increased three orders of magnitude since 2010 and will likely continue to increase in accordance with the number of wells on each chip. The sequencing cost is $5 per Mb and the read length ranges from 150 to 200 base pairs, which makes Ion Torrent less expensive and with read lengths shorter than 454, but more expensive and with read lengths slightly longer than Illumina and SOLiD. Ion Torrent reads have an error rate slightly higher than 1%, primarily caused by homopolymer sequences. These sequences are large stretches of the same nucleotide that cause a large release of H+, and above a certain

concentration, the quantitatively measured pH change becomes ambiguous, inadvertently causing an insertion or deletion. As the length of the homopolymer increases, the deletion error rate increases and the insertion error rate stays relatively constant (Ross et al. 2013), but Ion Torrent's short read length reduces the number of subsequent nucleotides in a read affected by these errors. The increasing daily throughput has recently expanded the use of Ion Torrent to animal cell sequencing experiments, including SNP discovery using *Bishu* mouse mutants (Damerla et al. 2014) and exomes using human samples (Boland et al. 2013).

A.4.2.2 Third-Generation Sequencing (TGS) Technologies

All four outlined SGS technologies are capable of sequencing a large number of nucleotides for modest time and financial investments when compared to Sanger sequencing, but all of the technologies require amplification. Large sequence copy numbers can lead to amplification-induced coverage and error biases (Keohavong and Thilly 1989) and this common attribute restricts certain characteristic ranges, despite the differences between the SGS technologies. Therefore, a fundamentally different group of sequencing technologies has emerged that address these limitations by avoiding the use of amplification. The TGS technologies covered here are two prevalent single molecule sequencing technologies and these TGS technologies' characteristics are displayed in Table A.2.

Table A.2: TGS Technology Characteristics

Common Name of TGS Technology	Sequencer	Read Length (kb per read)	Cost ($ per Mb)	Daily Throughput (Gb per day)	Run Time (hours per run)	Reads (million per run)
PacBio[a]	PacBio RS II	5.5-8.5[d]	50-150	0.100-0.375	0.5-3	0.05[i]
Nanopore[b]	Nanopore sequencing GridION[c]	10-30[e]	N/A[f]	10-50[g]	2-48+[h]	2[j]

[a]Website with sequencing data: http://www.pacificbiosciences.com/products/smrt-technology/
[b]Website with sequencing data: https://www.nanoporetech.com/technology/analytes-and-applications-dna-rna-proteins/dna-an-introduction-to-nanopore-sequencing
[c]1st generation
[d]Average value, but maximum ranges from 24-30 kb, depending upon the chemistry used
[e]Actual value is "tens of kb"
[f]The cost is not calculated on a per nucleotide basis, but on a per cartridge basis
[g]Actual value is "tens of Gb"
[h]Varies from minutes to days, depending on the experiment
[i]Per cell
[j]Calculated from the read length, daily throughput, and run time, but the actual number of reads depends on what molecule is sequenced and the length of time an experiment is run

A.4.2.2.1 PacBio

Pacific Biosciences' (PacBio) Single Molecule Real Time (SMRT) sequencing uses a modified, immobilized polymerase and fluorescence detection to accomplish exactly what the name states – single molecule real time sequencing. PacBio SMRT sequencing sample preparation consists of embedding individual DNA template strands into the bottom of a SMRT cell in 50 nm wide wells, called zero-mode waveguides (ZMW), along with an immobilized, modified DNA polymerase enzyme (Korlach et al. 2010, McCarthy 2010). A γ-phosphate-modified nucleotide solution is added to the SMRT cell. As the correct nucleotide is incorporated, the γ-phosphate

fluorophore causes a distinct, measurable fluorescent pulse. After the γ-phosphate tag is cleaved, the subsequent nucleotide can be added and sequencing continues.

PacBio SMRT sequencing reads are much longer than the SGS technology reads, averaging between 5.5 and 8.5 kilobases (kb). Because amplification is not required, amplification-based errors are eliminated, but there are multiple other error sources. The most common errors are insertions and deletions (Carneiro et al. 2012, Ross et al. 2013) and the average error rate is approximately 15% (Carneiro et al. 2012), resulting primarily from insertions within C-G rich regions. Compared to the SGS technologies, this is a larger error rate, but the stochastic nature of the error profile allows the standard Bayesian variant calling algorithm used by the Genome Analysis Toolkit to make robust nucleotide identifications (Carneiro et al. 2012). The error rate can be reduced if the sequence is short enough to permit repeat measurements during the same run, but repeat sequencing for large genomes can be difficult to accomplish. In addition to long read lengths, 50,000 reads are routinely produced per run, where the run time is between 30 minutes and 3 hours. The resulting average daily throughput is 200 Mb and the price per Mb ranges between $50 and $150. While both these characteristic values are inferior compared to the SGS technologies, the reduced throughput and higher cost are often considered a reasonable tradeoff for the extremely long read lengths. Full-length human RNA molecules with intron structures were sequenced using PacBio SMRT sequencing (Sharon et al. 2013), demonstrating its successful application towards a complex genome. While *de novo* genome assembly solely using PacBio data may not currently be fiscally optimal for large genomes because of the increased cost relative to the SGS methods, this technology can be used in combination with SGS technologies to aid in organizing the

reads for genome assembly (Koren et al. 2012). Other potential PacBio SMRT applications for the animal cell technology community include transcriptomic studies, SNP discovery, and epigenomic studies. PacBio sequencing technology produces long genomic reads and can simultaneously identify epigenetic modifications without additional reagents, sequencing preparations, or bisulfite conversion (Flusberg et al. 2010). Epigenomic studies are of particular interest because epigenetic modifications affect gene expression, cell differentiation, and other cell functions. Findings from these studies may include novel correlations between biomarkers and cell characteristics, including cellular productivity and cell age.

A.4.2.2.2 Nanopore Sequencing

Oxford Nanopore's Nanopore Sequencing was the first single molecule electrochemical sequencing technique (Clarke et al. 2009) not requiring fluorescently-labeled nucleotides; however, unlike Ion Torrent, the electrochemical measurement is induced by an electrical current rather than a pH change. Nanopore Sequencing's technology platform uses the protein staphylococcal α-hemolysin as a nanopore, where the hollow tube that is the α-hemolysin core, measuring a few nanometers in diameter, is situated in a synthetic polymer membrane on an array chip. This membrane has a high electrical resistance and the application of a potential across the membrane causes a current to flow through the nanopore's aperture. After a DNA-enzyme complex approaches the nanopore, the DNA strand is unzipped and each nucleotide individually passes through the nanopore (Rhee and Burns 2006). The nucleotide in the nanopore causes a distinguishable current disruption, permitting identification of the nucleotide sequence upon measurement of the changes in current.

Nanopore Sequencing read lengths have been reported in the tens of kb range, which are longer than the PacBio and SGS technology read lengths. Each array chip has many microwells and the number of wells used in parallel causes the run time to range from hours to days. The number of reads produced per run is approximately two million and the estimated daily throughput is in the tens of Gb, two orders of magnitude larger than PacBio's throughput. In addition to not being susceptible to amplification-induced biases, Nanopore Sequencing is not prone to fluorescence detection sensitivity-related error biases. However, uncertainty exists between the adenine and thymine electrochemical measurements, resulting in an initial sequencing error rate ranging from 1% to 10% (Clarke et al. 2009). The inability to differentiate between these nucleotides results in substitution errors and with a daily throughput of tens of Gb, this is a large number of incorrect nucleotides. The technology platform behind Nanopore Sequencing has been successfully demonstrated (Kowalczyk et al. 2012), yet there have not been reports of its application to animal cell culture sequencing. Nanopore Sequencing technology is also capable of sequencing molecules other than DNA, including RNA and proteins. Possible biomanufacturing applications of Nanopore Sequencing technology include epigenetic studies (Reinders et al. 2010), structural variation identification, protein identification, and metabolic small molecule detection (Clarke et al. 2009). Epigenetic studies would be practical because Nanopore Sequencing, like the PacBio SMRT sequencing technology, can measure epigenetic modifications without additional preparations, as methylated cytosine can be distinguished from the four standard DNA bases by its unique pore current value (Clarke et al. 2009). However, methylated cytosine's pore current value is between the

adenine and thymine base pore current values, making it more difficult to distinguish between those DNA bases, thus increasing the error rate.

A.5 Summary of Sequencing Technologies

The read length, cost, daily throughput, and error rate characteristics vary widely amongst the sequencing technologies, with the values spanning three to seven orders of magnitude for each characteristic. These differences are evident in Tables A.1 and A.2, but underlying characteristic correlations exist amongst the next-generation technologies and these correlations are not well portrayed through the tabular displays. Figure A.1 depicts these correlations by plotting all the technologies' values for two of the characteristics on one graph. When plotted in this way, the illustration reveals how each sequencing technology occupies an exclusive niche within the sequencing technology community and illustrates the underlying correlations. Briefly, an increased daily throughput correlates with both a decreasing read length and a decreasing cost, as illustrated in Figures A.1(a) and A.1(b), respectively. It can be concluded from these correlations that as the read length increases amongst the technologies, the sequencing cost also increases (Figure A.1(c)). Correlations also exist between these three characteristics and the error rate. As the error rate increases amongst the sequencing technologies, so does the read length and the sequencing costs, while daily throughput decreases (Figures A.1(d), A.1(e), and A.1(f), respectively). The correlations amongst the SGS and TGS technologies are quite strong and while Sanger aligns with some of these trends, the differences between Sanger and the next-generation technologies highlight the advancements in sequencing technology. Understanding each technology's characteristics in relation to

other technologies enables the identification of which technology is best suited for an application.

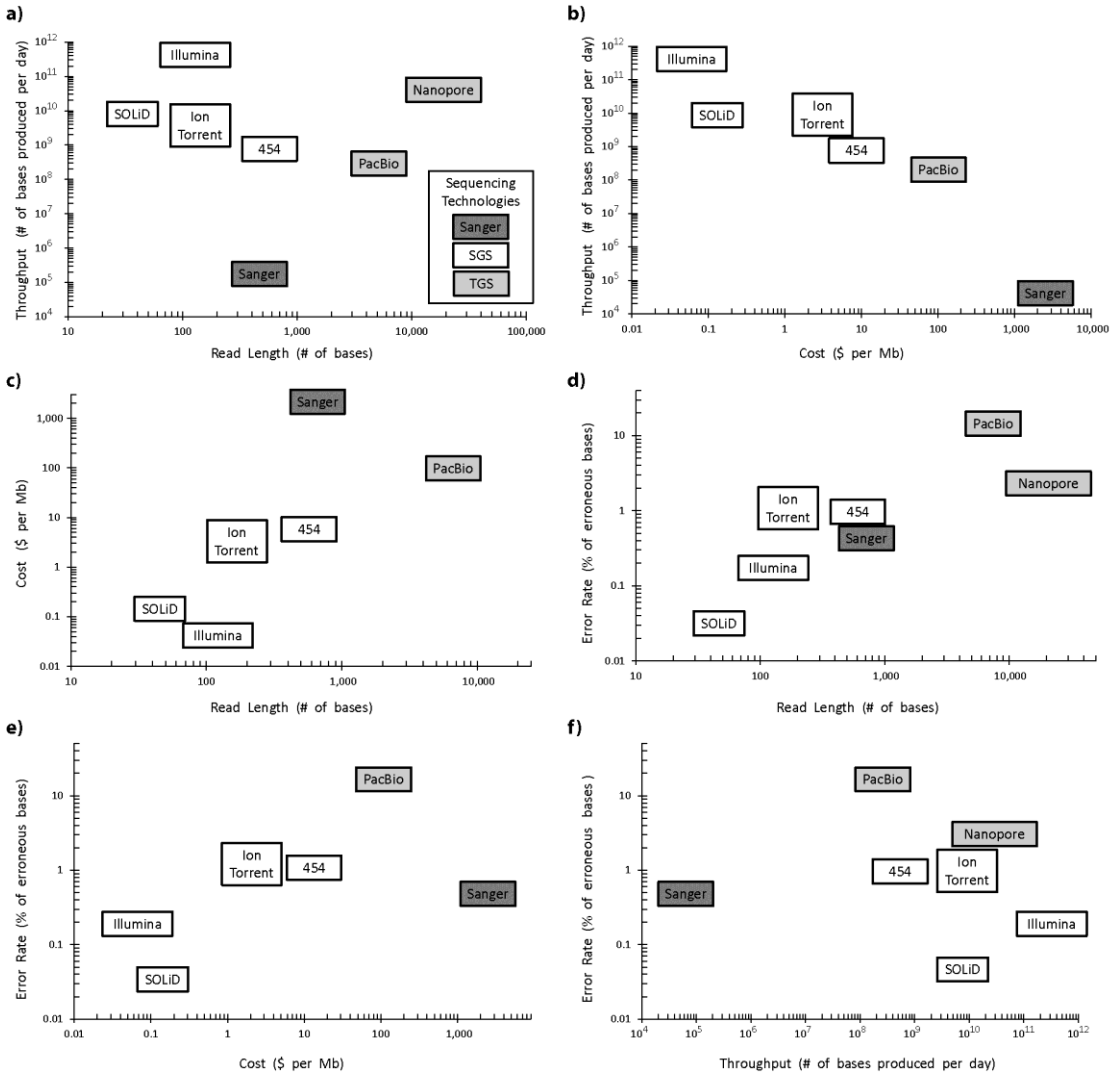


Figure A.1: Comparison of the read length, cost, daily throughput, and error rate characteristics for the sequencing technologies. The panels compare a) throughput vs read length, b) throughput vs cost, c) cost vs read length, d) error rate vs read length, e) error rate vs cost, and f) error rate vs throughput. Note that both axes use logarithmic scales and Nanopore Sequencing is not represented in sub-plots b), c), and e) because the cost is not yet available.

A.6 Applications of Sequencing Technologies

Advances in sequencing technologies have resulted in an increasing average read length, decreasing sequencing cost per base, and more selective chemistries. Applications previously deemed too expensive, labor intensive, or complex may soon be feasible (Becker et al. 2011). Despite this potential, many projects encounter issues resulting from each technology's specific advantages and disadvantages. For example, Illumina assemblies have a large number of reads, but the reads are short and difficult to accurately assemble. Alternatively, PacBio assemblies have long reads, but require a significantly higher cost and potentially have a higher error rate. Projects increasingly rely on using multiple technologies in parallel to minimize technology-specific disadvantages. Genome assemblies using this approach are commonly referred to as hybrid assemblies and with the appropriate selection of complementary sequencing technologies, significant advancements are possible. However, the sequence assembly may not be improved if the selected technologies do not have complementary characteristics that result in reduced bias or increased coverage. Technologies located further from each another in Figure A.1 tend to be more complementary. For example, a hybrid assembly using PacBio and Illumina technologies exploits Illumina's low-cost characteristic to generate the majority of the sequence reads, but relies on PacBio's long reads to aid in organizing a complete assembly. This specific hybrid assembly has been successfully demonstrated with parrot and bacterial genomes (Koren et al. 2012). Additional mammalian hybrid sequencing experiments have been performed using Sanger and PacBio sequencing technologies for human embryonic stem cell transcriptome sequencing (Au et al.

2013) and using 454 and Illumina technologies for *Arctocephalus gazelle* SNP genotyping (Hoffman et al. 2012) and *Artibeus jamaicensis* transcriptome sequencing (Shaw et al. 2012). As sequencing technologies continue to advance, hybrid assemblies will likely become more commonplace within the animal cell culture community.

Bioprocess-related applications of sequencing technologies within the animal cell technology community are increasing in number and sophistication. One application with great potential for expanded use is the sequencing of single cells, which can generate a cell-specific foundation for quantitative cell characterization and biological understanding. Comparative analysis across many individual cells may identify gene-regulatory network patterns and reveal differences in the transcribed genes between individual cells and populations, providing additional biological insight into cellular processes and functional states. High-resolution transcriptional maps could potentially be assembled from these data and used to assess the transcriptional regulation of important cellular processes by identifying transcript expression levels and variants. This information could be used to identify biomarkers of relevant cell phenotypes, such as those associated with genetic instability or unexpected changes in viability and aid in cellular engineering efforts by identifying potentially beneficial functional genomic modulations. As the biopharmaceutical industry transitions to continuous biomanufacturing campaigns, the effects of long-term cell culture and genomic instability will need to be evaluated. Continuous monitoring of quantitative expression profiles from genome and transcript sequencing can be used to characterize cell age-related variations in host cell protein profiles, cellular phenotypes, and production-related traits. These applications may facilitate improved processes with

greater process control linked to changes previously identified from single cell analysis.

A.7 Conclusions

For years, the sequencing technology benchmark has been to sequence a complete human genome for less than $1,000 (Service 2006). This arbitrary threshold value represents the time point when the cost of personal genomics within the medical community is no longer fiscally unreasonable. At the JP Morgan Healthcare Conference in early 2014, Illumina announced this threshold had been achieved (Bio-IT 2014) as their newest sequencer, the HiSeq X, is capable of sequencing a complete human genome for $998. This total accounts for the components that constitute a complete sequencing experiment, including the core consumables, sequencer depreciation cost, and sample preparation and labor (but not assembly nor annotation). The HiSeq X sequencer uses the SBS technology platform with a redesigned flow cell that's equipped with patterned nanowells, a chemistry that is four times faster, and uses bidirectional optical scanning, accounting for a six-fold improvement in speed (Krol 2014). In addition to achieving the $1,000 goal, these improvements reduce the standard Illumina experiment sequencing time from twenty-three hours to five hours, greatly increasing the throughput. Considering the first human genome cost billions of dollars and required more than ten years to sequence (Lander et al. 2001, Venter et al. 2001), the advancement is clear. As technology continues to advance and sequencing technologies become cheaper, faster, and more diverse, the application of sequencing technologies towards animal cell research and the 21st century's medical challenges will continue to become increasingly relevant.

A.8 Acknowledgements

I am grateful for support from the National Science Foundation under grant no. 1124647 and the National Institute of Standards and Technology under grant no. 60NANB11D185.

REFERENCES

Ansorge WJ. (2009) Next-generation DNA sequencing techniques. New Biotechnol. 25:195-203.

Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA, Bakel Hv, Schadt EE, Reijo-Pera RA Underwood JG, Wong WH. (2013) Characterization of the human ESC transcriptome by hybrid sequencing. P Natl Acad Sci USA. 110:E4821-E4830.

Baker M. (2010) Nanotechnology imaging probes: smaller and more stable. Nat Methods. 7:957-962.

Becker J, Hackl M, Rupp O, Jakobi T, Schneider J, Szczepanowski R, Bekel T, Borth N, Goesmann A, Grillari J, Kaltschmidt C, Noll T, Puhler A, Tauch A, Brinkrolf K. (2011) Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. J Biotechnol. 156:227-235.

Bennett S. (2004) Solexa ltd. Pharmacogenomics. 5:433-438.

BioIT World. (2014) Illumina announces the thousand dollar genome. Bio-IT World.

Boland JF, Chung CC, Roberson D, Mitchell J, Zhang X, Im KM, He J, Chanock SJ, Yeager M, Dean M. (2013) The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing. Hum Genet. 132:1153-1163.

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

Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. (2012) Pacific biosciences sequencing technology for genotyping and variation and discovery in human data. BMC Genomics. 13:375-382.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 4:265-270.

Cervera L, Gutierrez-Granados S, Martinez M, Blanco J, Godia F, Segura MM. (2013) Generation of HIV-1 Gag VLPs by transient transfection of HEK 293 suspension cell culture using an optimized animal-derived component free medium. J Biotechnol. 166:152-165.

Damerla RR, Chatterjee B, Li Y, Francis RJB, Fatakia SN, Lo CW. (2014) Ion Torrent sequencing for conducting genome-wide scans for mutation mapping analysis. Mamm Genome. 25:120-128.

Everett MV, Grau ED, Seeb JE. (2011) Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. Mol Ecol Resour. 11:93-108.

Flosberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 7:461-465.

Galan M, Guivier E, Caraux, Charbonnel N, Cosson JF. (2010) A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. BMC Genomics. 11:296.

Glenn TC. (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour. 11:759-769.

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

Hammond S, Swanberg FC, Kaplarevic M, Lee KH. (2011) Genomic sequencing and analysis of a Chinese hamster ovary cell line using Illumina sequencing technology. BMC Genomics. 12:67.

Hoffman JI, Tucker R, Bridgett SF, Clark MS, Forcada J, Slate J. (2012) Rates of assay success and genotyping error when single nucleotide polymorphism genotyping in non-model organisms: a case study in the Antarctic fur seal. Mol Ecol Resour. 12:861-872.

Johnson KC, Yongky A, Vishwanathan N, Jacob NM, Jayapal KP, Goudar CT, Karypis G, Hu WS. (2014) Exploring the transcriptome space of a recombinant BHK cell line through next generation sequencing. Biotechnol Bioeng. 111:770-781.

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

Keohavong P, Thilly WG. (1989) Fidelity of DNA polymerases in DNA amplification. P Natl Acad Sci USA. 86:9253-9257.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Phillippy AM. (2012) Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat Biotechnol. 30:693-700.

Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW. (2010) Real-time DNA sequencing from single molecule polymerase molecules. Method Enzymol. 472:431-455.

Kowalczyk SW, Wells DB, Aksimentiev A, Dekker C. (2012) Slowing down DNA translocation through a nanopore in lithium chloride. Nano Lett. 12:1038-1044.

Kremkow B, Lee KH. (2013) Next-generation sequencing technologies and their potential impact on CHO cell-based biomanufacturing. Pharm Bioproc. 1:455-465.

Krol A. (2014) What you need to know about Illumina's new sequencers. Bio-IT World

La Merie Business Intelligence. (2013) Blockbuster biologics 2012. R&D Pipeline News. 7:2-28.

Lander ES, Linton LM, Birren B, *et al.* (2001) Initial sequencing and analysis of the human genome. Nature. 409:860-921.

Lemay MA, Henry P, Lamb CT, Robson KM, Russello MA. (2013) Novel genomic resources for a climate change sensitive mammal: characterization of the American pika transcriptome. BMC Genomics. 14:311.

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

Liu L, Li YH, Li SL, Hu N, He Y, Pong R, Lin D, Lu L, Law M. (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012:1-11.

Mardis ER. (2008) Next-generation DNA sequencing methods. Annu Rev Genom Hum G. 9:387-402.

McCarthy A. (2010) Third generation DNA sequencing: Pacific Biosciences' Single Molecule Real Time technology. Chem Biol. 17:675-676.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297-1303.

Merriman B, Rothberg JM, Ion Torrent R&D Team. (2012) Progress in Ion Torrent semiconductor chip based sequencing. Electrophoresis. 33:3397-3417.

Metzker ML. (2010) Sequencing technologies – the next generation. Nat Rev. 11: 31-46.

Naik AD, Menegatti S, Gurgel PV, Carbonell RG. (2011) Performance of hexamer peptide ligands for affinity purification immunoglobulin G from commercial cell culture media. J Chromatogr A. 1218:1691-1700.

Picotti P, Aebersold R. (2012) Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods. 9:555-566.

Reinders J, Pasazkowski J. (2010) Bisulfite methylation profiling of large genomes. Epigenomics. 2:209-220.

Rhee M, Burns MA. (2006) Nanopore sequencing technology: research trends and applications. Trends Biotechnol. 24:580-586.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. (2013) Characterizing and measuring bias in sequence data. Genome Biol. 14:51-70.

Rothberg JM, Leamon JH. (2008) The development and impact of 454 sequencing. Nat Biotechnol. 26:1117-1124.

Sanger F, Nicklen S, Coulson AR. (1977) DNA sequencing with chain-terminating inhibitors. P Natl Acad Sci USA. 74:5463-5467.

Schadt EE, Turner S, Kasarskis A. (2010) A window into third-generation sequencing. Hum Mol Genet. 19:R227-R240.

Schuster SC. (2008) Next-generation sequencing transforms today's biology. Nat Methods. 5:16-18.

Scherer WF, Syverton JT, Gey GO. (1953) Studies on the propagation in vitro of poliomyelitis viruses 4 – viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix. J Exp Med. 97:695-710.

Service RF. (2006) Gene sequencing - The race for the $1000 genome. Science. 311: 1544-1546.

Sharon D, Tilgner H, Grubert F, Snyder M. (2013) A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 31:1009-1014.

Shaw TI, Srivastava A, Chou WC, Liu L, Hawkinson A, Glenn TC, Adams R, Schountz T. (2012) Transcriptome sequencing and annotation for the Jamaican fruit bat (*Artibeus jamaicensis*). Plos One. 7:1-12.

Shendure J, Ji H. (2008) Next-generation DNA sequencing. Nat Biotechnol. 26:1135-1145.

Venter JC, Adams MD, Myers EW, *et al.* (2001) The sequence of the human genome. Science. 291:1304-1351.

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

Yu M, Hu Z, Pacis E, Vijayasankaran N, Shen A, Li F. (2011) Understanding the intracellular effect of enhanced nutrient feeding toward high titer antibody production process. Biotechnol Bioeng. 108:1078-1088.

Appendix B

DREAM-ZYP AND GLYCO-MAPPER DEVELOPMENT

B.1 Preface

This appendix explores the detailed creation of the novel Discretized Reaction Network Modeling using Fuzzy Parameters (DReaM-zyP) technique and its application towards CHO glycosylation through the creation of Glyco-Mapper. The DReaM-zyP technique combines portions of three common modeling techniques: genome-scale reconstruction, kinetic modeling, and fuzzy logic modeling. Each of these techniques will be briefly explained and the components of each that were used for DReaM-zyP will be detailed. The glycoform images used to represent the experimental results and modeling predictions are explained.

B.2 Materials and Methods

B.2.1 Glycosylation Reaction Network Genes in the CHO and CH Genomes

The Chinese hamster (CH), identified as *Cricetulus griseus,* is the origin of the Chinese hamster ovary (CHO) cell, also referred to as *Cricetulus griseus* within the National Center for Biotechnology Information (NCBI) repository (National Center

for Biotechnology Information 2013). The annotated CHO-K1 genome was published in 2012 (Xu et al. 2012) and the annotated CH genome was published in 2014 (Lewis et al. 2014), at which time the CHO-K1 genome sequence was re-annotated. The Glyco-Mapper CHO glycosylation genes were primarily obtained from Xu et al. Supplementary Table 13, metabolism genes were obtained from literature (Ahn and Antoniewicz 2012), and additional confirmed genes were supplemented (Bosques et al. 2010). The gene sequences for the 59 N-glycosylation and 92 metabolism-related genes associated with glycosylation were manually curated and the corresponding enzymatic reaction information was cataloged. Available gene sequences were obtained for each of the three *Cricetulus griseus* NCBI annotations publicly available at CHOgenome.org (Hammond et al. 2012; Kremkow et al. 2015).

B.3 Results and Discussion

B.3.1 DReaM-zyP Technique Influences

B.3.1.1 Genome-Scale Reconstruction

Genome-scale network reconstruction has become an important tool in the study of systems biology of metabolism (Thiele et al. 2005; Covert et al. 2004; Feist and Palsson 2008). Reconstructions require genomic data from the target organism to enable the creation of species-specific databases. The general steps to a genome-scale reconstruction involve creation of a draft reconstruction, manual refinement of the

reconstruction, conversion of the reconstruction to a mathematical model, network evaluation, and utilization (Thiele and Palsson 2010). The mathematical models are able to predict important physiological system properties. This modeling technique is quite labor and time intensive, but creates species-specific mathematical models capable of predicting important cellular characteristics and is often applied specifically to metabolism (Duarte et al. 2007; Thiele and Palsson 2010).

The draft reconstruction of the CHO glycosylation reaction network and related central carbon metabolism (CCM) and nucleotide sugar transport metabolic pathways was created using the CHO-K1 and CH genome annotations. As per the protocol (Thiele and Palsson 2010), substrate and cofactor usage, neutral enzymatic reactions, gene and reaction localization, heteromeric enzyme complexes, isozyme functionalities, intracellular transport mechanisms, and supporting metabolic reactions were all verified. The manual reconstruction and refinement resulted in comprehensive reaction databases for candidate glycosylation and metabolic functions (Tables F.1 and F.2). The conversion of the reconstruction to the mathematical model deviated from the genome-scale reconstruction protocol, as the metabolism functions and parameters were not applicable towards the glycosylation objectives.

B.3.1.2 Kinetic Modeling

Kinetic enzymatic reaction modeling is more than a century old (English et al. 2006) and has been extensively studied. Many enzyme kinetic equations use assumptions to achieve mathematical solutions and complex kinetics account for allosteric, immobilized, and inhibited enzyme reactions while incorporating the effect of the temperature and pH. Kinetic equations have been used in many glycosylation

models (Umaña and Bailey 1997, Krambeck and Betenbaugh 2005, Krambeck et al. 2009) to achieve successful analysis of glycosylation with increasing complexity as the models improve. The enzymes responsible for the glycan substrate modifications are critically important to the final glycoform, especially as glycosylation is a non-template driven process. DReaM-zyP incorporated kinetics into Glyco-Mapper by requiring one kinetic activity level value (k_{ALV}) parameter (combination of gene expression and kinetic activity values) for each gene. This assumption enables kinetic-based glycoform predictions while simplifying the mathematics.

B.3.1.3 Fuzzy Logic Modeling

Fuzzy logic modeling is traditionally used to transform quantitative values into qualitative (discrete number) descriptors using a predetermined set of heuristic rules to identify interactions (Sokhansanj et al. 2009). Numerical data is converted to discrete values enabling data cluster identification (Zhang et al. 2009). The most common application of fuzzy logic modeling uses gene expression data to identify uncharacterized protein and transcription factor functions or interactions (Ressom et al. 2005). This method attempts to model multi-scale biomolecular network models by using a large amount of data generated from a small number of experiments (Sokhansanj et al. 2009). Glycosylation is a multi-scale biomolecular reaction network and while there is not a massive amount of data, glycosylation data is measurable. DReaM-zyP incorporated fuzzy logic modeling into Glyco-Mapper by discretizing the gene k_{ALV} parameters to further simplify the reaction kinetics. This is justified due to the complexity of obtaining accurate gene expression and enzyme activity values. The objective function is not the identification of interacting gene clusters, but the

identification of glycoform patterns because the glycosylation reaction network has previously been determined.

B.3.2 Glyco-Mapper Creation Using DReaM-zyP

A CHO-specific gene reconstruction of the CHO glycosylation reaction network, including the glycosylation-related CCM network and nucleotide sugar transport reactions using the CHO-K1 and CH genomes as the first step of DReaM-zyP. The mathematical conversion of the reconstruction incorporated fuzzified kinetic reaction parameters to produce glycoform predictions and Glyco-Mapper was created. Similar to genome-scale reconstructions, Glyco-Mapper is iterative and the k_{ALV} parameters should be iteratively optimized to represent the experimental reference glycoform. After establishment of the reference k_{ALV} parameter settings, parameter adjustments representing any media feed alteration or glycosylation, metabolism, or transport gene overexpression, knockdown, or knockout generate the corresponding glycoform prediction. The Glyco-Mapper tool attempts to predict biopharmaceutically relevant glycoforms resulting from experimentally realistic alterations and measurements.

B.3.3 Glyco-Mapper Predicted Glycoforms

The reference and predicted glycoforms are currently predicted in an Excel spreadsheet. Visual representation of the prediction data (in a figure) enables improved trend identification, particularly when experimental data is included. Four separate glycoforms are able to be customized depending on the biotherapeutic and

glycoform location parameter settings. The biopharmaceutical parameter can be set to mAb or non-mAb and the glycoform location parameter can be set to secreted or intracellular. The parameter-defined glycoforms include [mAb - secreted] in Figure B.1, [non-mAb - secreted] in Figure B.2, [mAb - intracellular] in Figure B.3, and [mAb - intracellular] in Figure B.4.

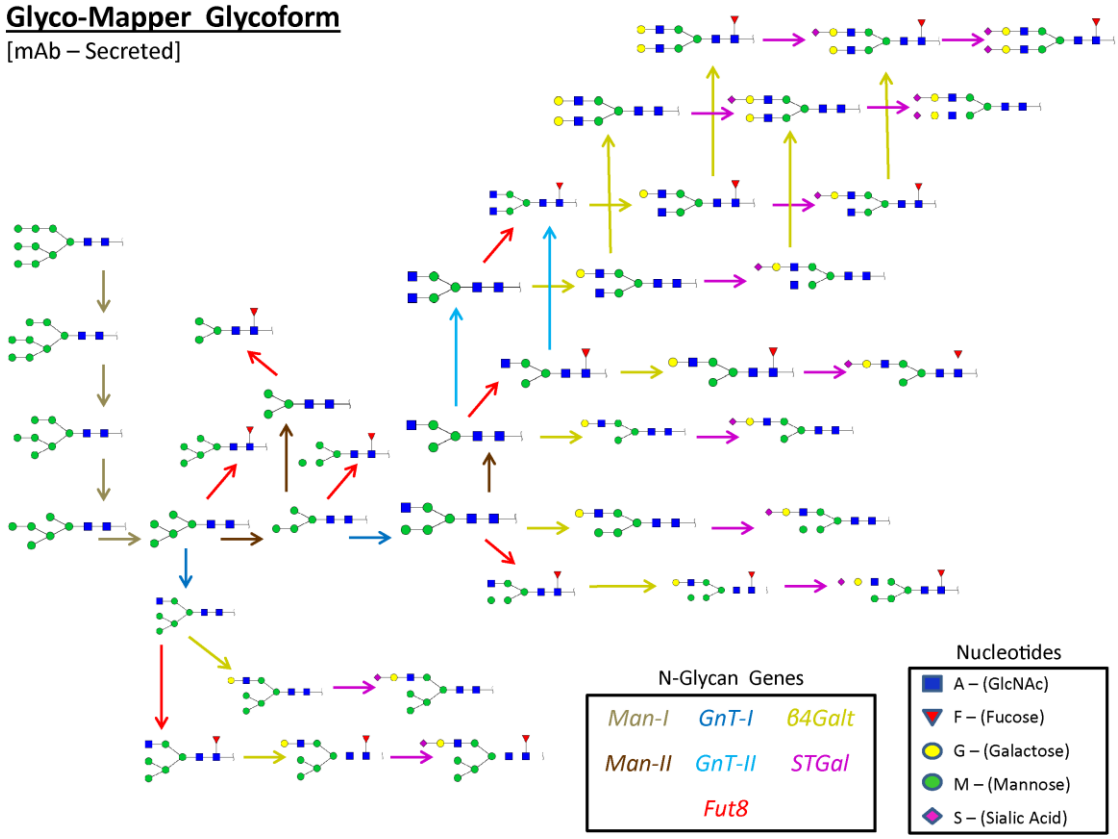


Figure B.1: The Glyco-Mapper [mAb - secreted] stock glycoform.

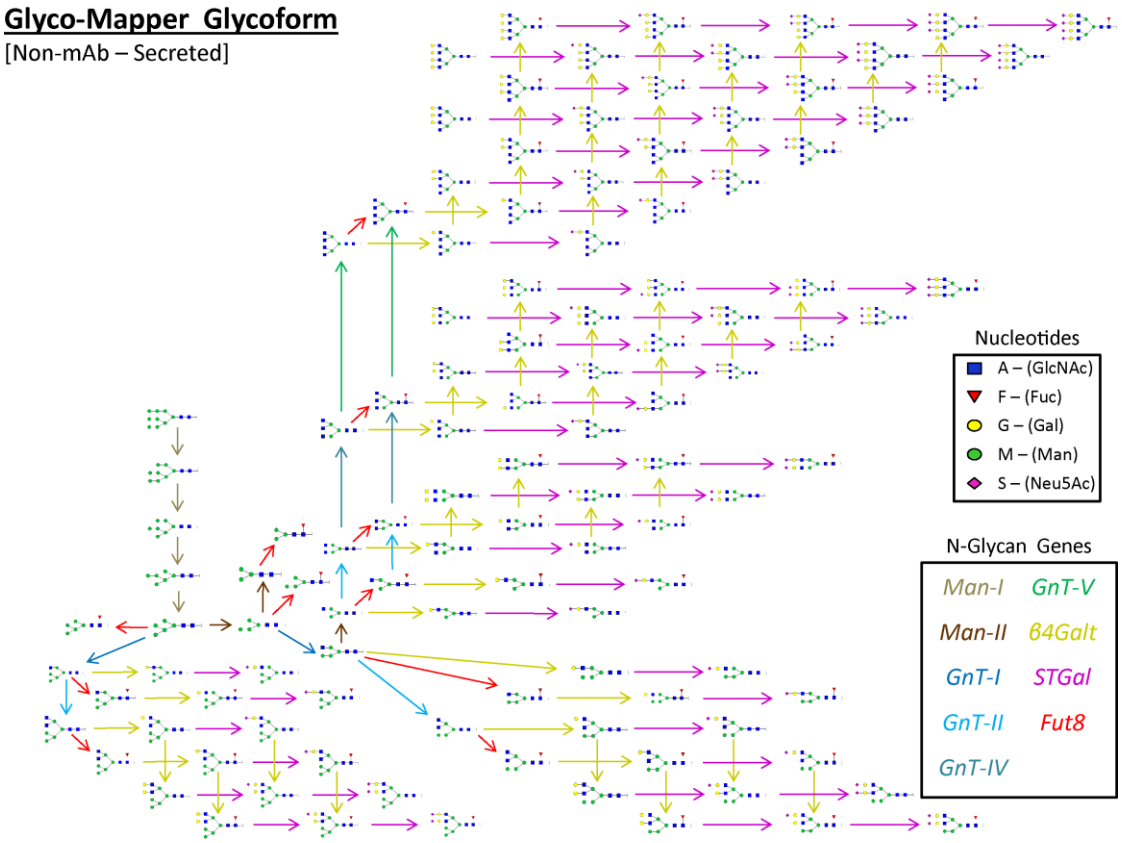


Figure B.2: The Glyco-Mapper [non-mAb - secreted] stock glycoform.

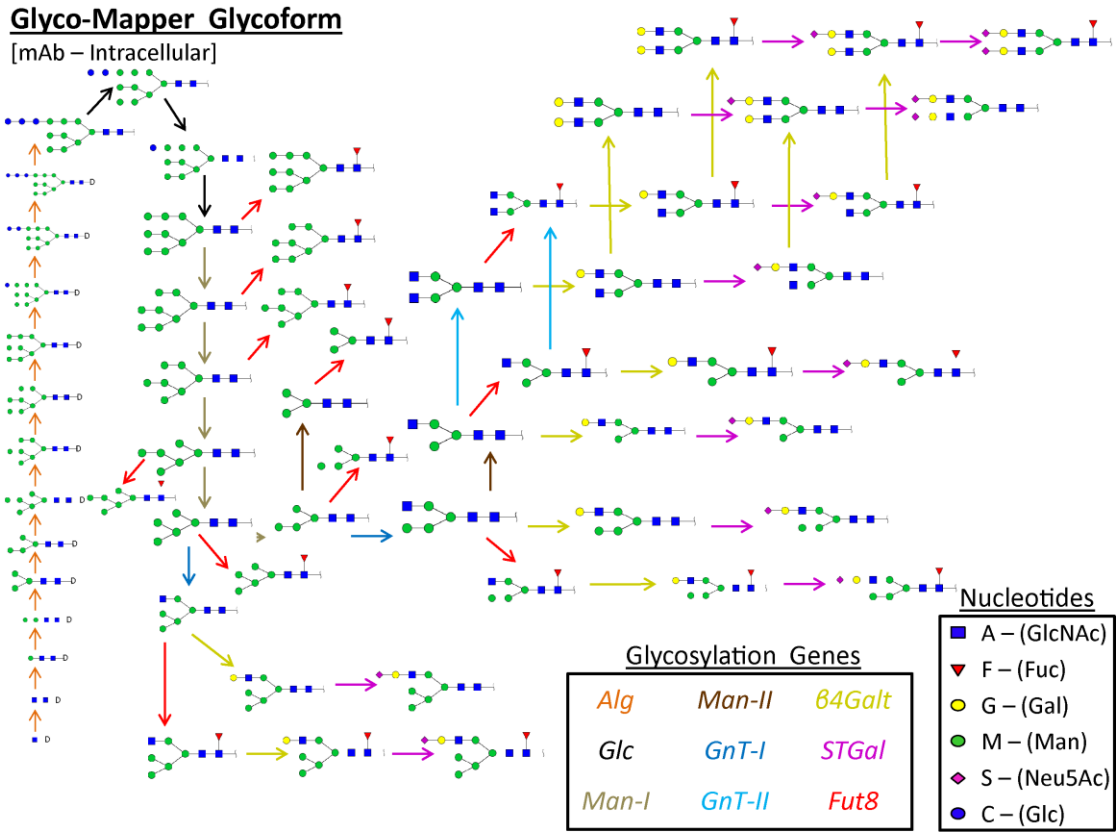


Figure B.3: The Glyco-Mapper [mAb - intracellular] stock glycoform.

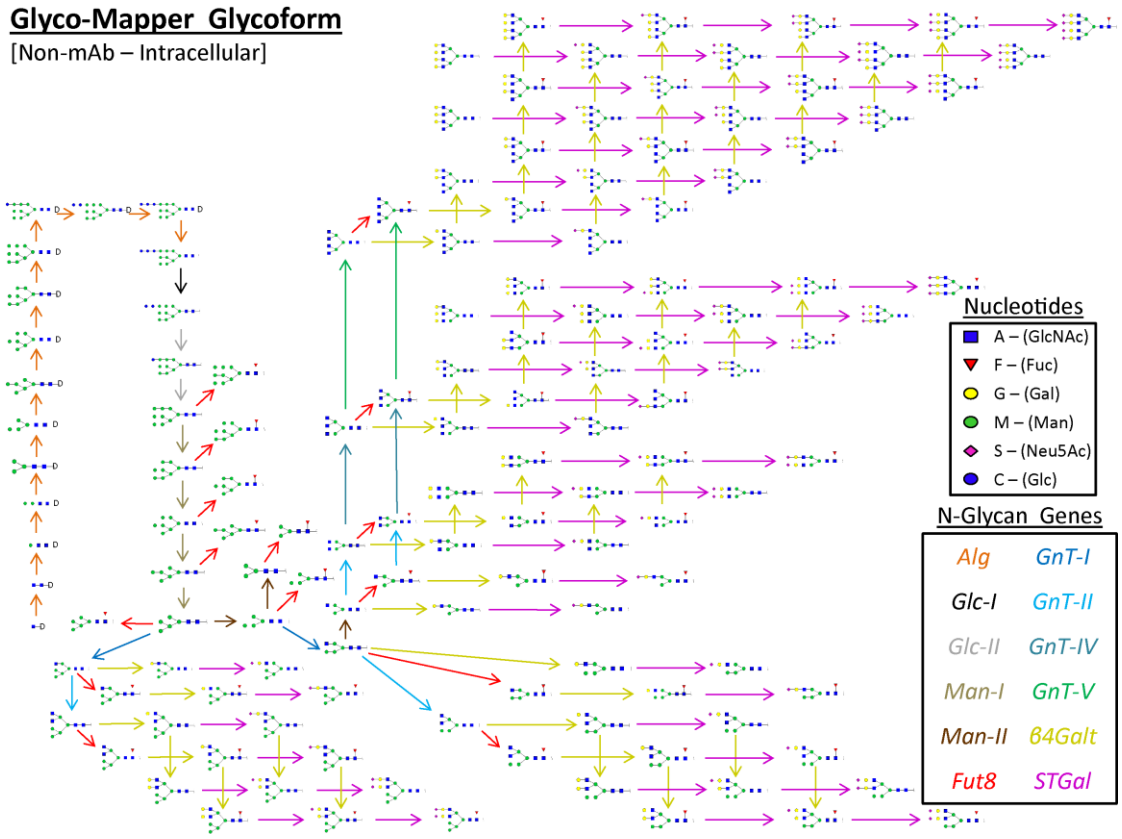


Figure B.4: The Glyco-Mapper [non-mAb - intracellular] stock glycoform.

REFERENCES

Ahn WS, Antoniewicz MR. (2012) Towards dynamic metabolic flux analysis in CHO cell cultures. Biotechnol J. 7:61-74.

Bosques CJ, Collins BE, Meador JW, Sarvaiya H, Murphy JL, DelloRusso G, Bulik DA, Hsu IH, Washburn N, Sipsey SF, Myette JR, Raman R, Shriver Z, Sasisekharan R, Venkataraman G. (2010) Chinese hamster ovary cells can produce galactose-α-1,3-galactose antigens on proteins. Nat Biotechnol. 28:1153-1156.

Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BØ. (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature. 429:92-96.

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. PNAS. 104:1777-1782.

English BP, Min W, van Oijen AM, Lee KT, Luo G, Sun H, Cherayil BJ, Kou SC, Xie XS. (2006) Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. Nat Chem Biol. 2:87-94.

Feist AM, Palsson BØ. (2013) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol. 26:659-667.

Hammond S, Kaplarevic M, Borth N, Betenbaugh MJ, Lee KH. (2012) Chinese hamster genome database: an online resource for the CHO community at www.CHOgenome.org. Biotechnol Bioeng. 109:1353-1356.

Krambeck FJ, Betenbaugh MJ. (2005) A Mathematical Model of N-Linked Glycosylation. Biotechnol Bioeng. 92:711-728.

Krambeck FJ, Bennum SV, Narang S, Choi S, Yarema KJ, Betenbaugh MJ. (2009) A mathematical model to derive N-glycan structures and cellular enzyme activities from mass spectrometric data. Glycobiology. 19:1163-1175.

Kremkow BG, Baik JY, MacDonald ML, Lee KH. (2015) CHOgenome.org 2.0: Genome resources and website updates. Biotechnol J. 10:931-938.

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

National Center for Biotechnology Information (US). (2013) The NCBI Handbook. National Center for Biotechnology Information (US), Bethesda.

Ressom H, Natarajan P, Varghese RS, Musavi MT. (2005) Applications of fuzzy logic in genomics. Fuzzy Set Syst. 152:125-138.

Sokhansanj BA, Datta S, Hu X. (2009) Scalable dynamic fuzzy biomolecular network models for large scale biology. Fuzzy Systems in Bio. 1:235-255.

Thiele I, Palsson BØ. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 5:93-121.

Thiele I, Price ND, Vo TD, Palsson BØ. (2005) Candidate metabolic network states in human mitochondria: impact of diabetes, ischemia and diet. J Biol Chem. 280:11683-11695.

Umaña P, Bailey JE. (1997) A Mathematical Model of N-linked Glycoform Biosynthesis. Biotechnol Bioeng. 55:890-908.

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

Zhang S, Wang RS, Zhang XS, Chen L. (2009) Fuzzy system methods in modeling gene expression and analyzing protein networks. Fuzzy Systems in Bio. 1:235-255.

Appendix C

SUPPLEMENTAL GLYCO-MAPPER PREDICTIONS

C.1 Preface

This appendix explores the additional Glyco-Mapper predictions of the CHO cell-engineering literature listed in Table 4.1, but not pictured or discussed in Chapter 4. These predictions include heterologous gene expression, gene overexpression, and gene knockouts affecting the glycoform sialylation, galactosylation, antennarity, and fucosylation using glycosylation and metabolism genes. Two uncommon biopharmaceutical glycan linkages consistent with published CHO glycoforms were included in the Glyco-Mapper databases and sample glycoforms containing these linkages are demonstrated. The incorrectly predicted glycans presented in Table 4.4 are further categorized and explained.

C.2 Materials and Methods

C.2.1 Glyco-Mapper Validation Procedure

The glycosylation and metabolism enzyme k_{ALV} parameters were first set to the default reference k_{ALV} parameter settings as a temporary reference. The summary

parameters [type of biopharmaceutical, glycoform of interest, media composition, glycan of interest] were then defined [mAb or non-mAb; secreted or internal; presence or absence of Glc, Fuc, GlcN, Gal, ManNAc, GalNAc, GlcA, SO4, S, TSO4; glycan of interest as defined in Glyco-Mapper] in accordance with the study of interest. The glycosylation and metabolism enzyme k_{ALV} parameters were adjusted to values that represented a reasonable guess for the reference glycoform. The predicted reference glycoform was analyzed and compared to the reference experimental glycoform. From this analysis, the gene k_{ALV} parameters that should be altered to improve the predicted reference glycoform were determine. The selected k_{ALV} parameters were altered, the predicted glycoform was analyzed and then compared again to the reference experimental glycoform. This process continued until the reference glycoform had been optimized.

Glycoform-altering strategies using glycosyltransferase, metabolism, or transporter gene or nutrient feed composition alterations can now be predicted. Using this optimized set of reference k_{ALV} parameter settings, the corresponding glycosylation, metabolism, or transport gene k_{ALV} parameters were altered to represent gene overexpression, knockdown, or knockout experiments. The media composition list was adjusted if any sugar feeds were introduced or removed and then the resulting predicted glycoform was analyzed.

C.3 Results

C.3.1 Confirmation of Glyco-Mapper Predictions Replicating Publication Data

Glyco-Mapper predicted 17 cell-altered glycoforms from 10 publications and 9 of the glycoforms from 5 of the publications were presented in Chapter 4. The 8 glycoforms from the remaining 5 publications are analyzed in this appendix. The Glyco-Mapper predictions presented here specifically involved engineered fucosylation, sialylation, galactosylation, antennarity, and metabolism (GDP-Fuc) alterations. Again, Glyco-Mapper achieved an overall 96.1% glycan prediction accuracy (1,547 of 1,608 glycans) and 85% delta accuracy, as well as an average glycoform prediction sensitivity and specificity of 85% and 97%, respectively.

C.3.1.1 Strategy 1: Expression of Heterologous Glycosyltransferases

In addition to the heterologous glycoform-engineering study by Onitsuka (Onitsuka et al. 2012), Glyco-Mapper successfully modeled and predicted the glycan distributions reported by Naso (Naso et al. 2010) where Naso expressed *SiaA* to decrease sialylation, thereby potentially decreasing the IgG's biotherapeutic *in vivo* half-life. Glyco-Mapper modeled the wild type (Figure C.1) with 38 of 40 correct glycans [5 of 7 present; 33 of 33 absent]. When *SiaA* expression was modeled (Figure C.2), Glyco-Mapper predicted 39 of 40 glycans correctly [3 of 4 present; 36 of 36 absent] and data confirmed each of the 3 glycans that changed prediction classes. The

predicted IgG glycoform resulting from the altered heterologous glycosyltransferase expression was accurate, sensitive, and specific.

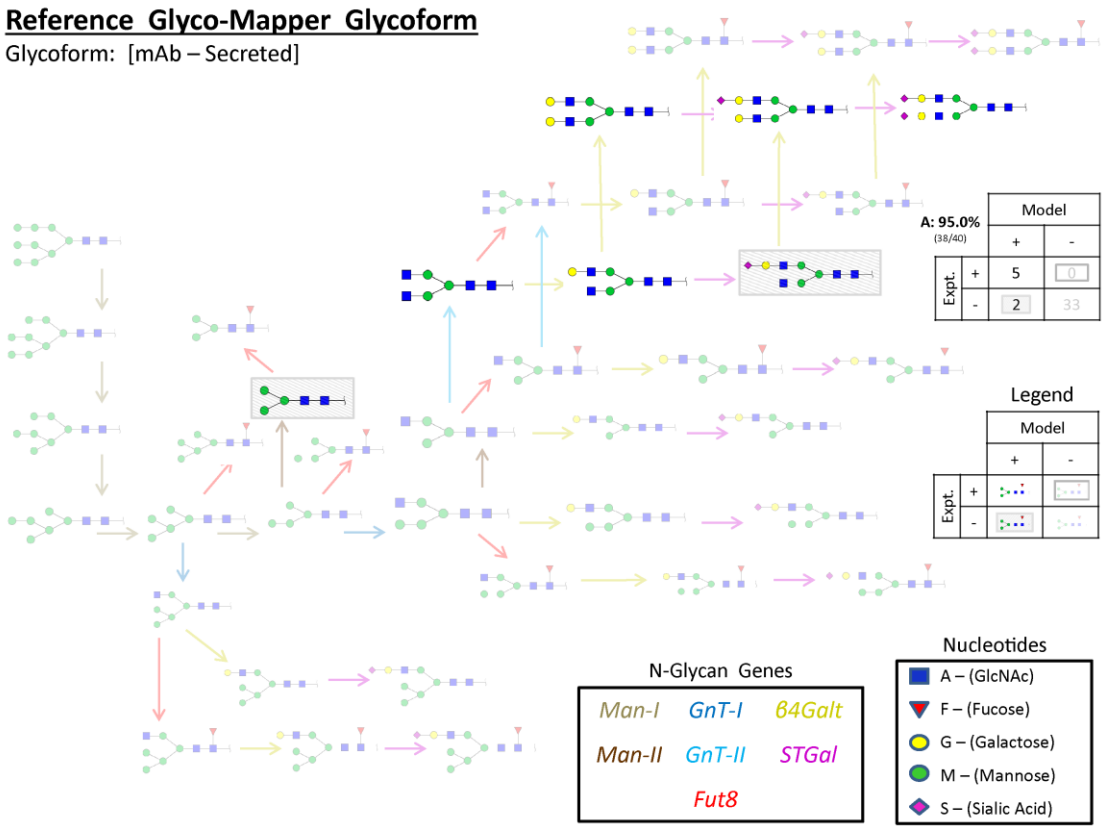


Figure C.1: The Glyco-Mapper prediction of the Naso *et al.* reference glycoform. The afucosylated, bi-antennary glycans A2, A2G1, A2G2, A2G2S1, and A2G2S2 are correctly predicted to be experimentally present in this reference mAb glycoform.

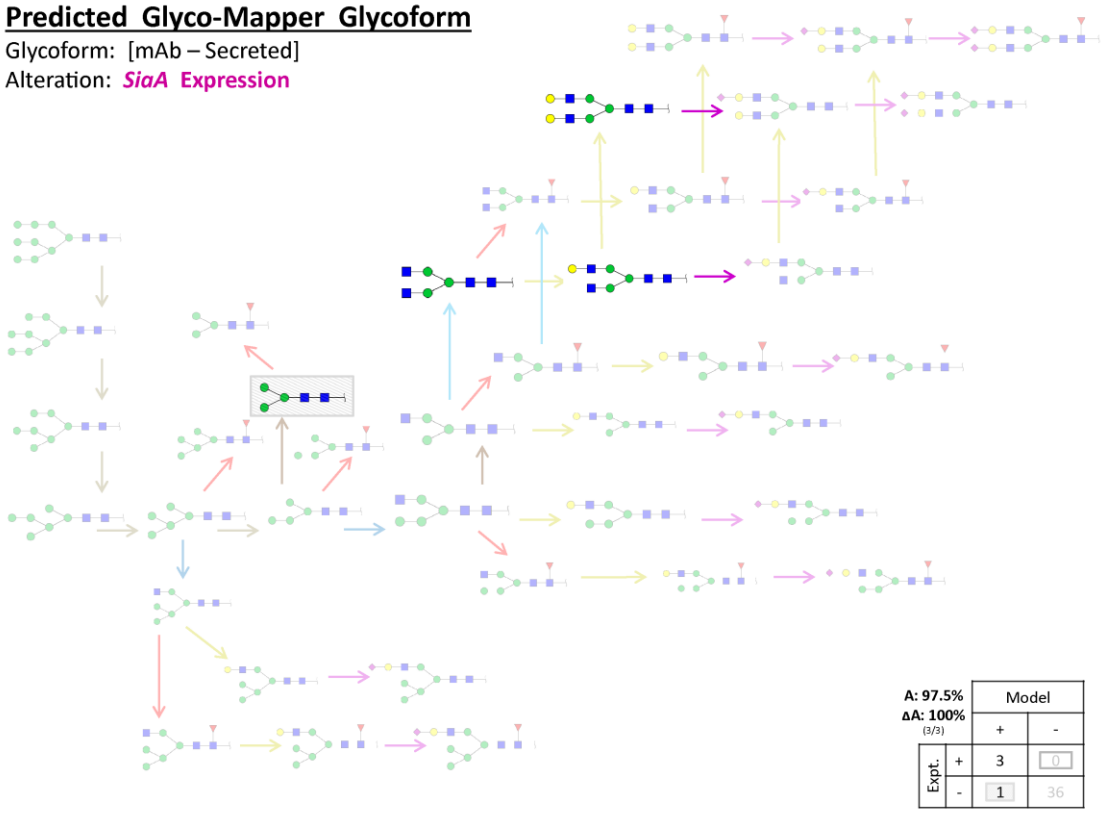


Figure C.2: The Glyco-Mapper prediction of the *SiaA* expression based on the Naso *et al.* reference glycoform (Figure C.1). The asialylated glycans A2, A2G1, and A2G2 are correctly predicted to be experimentally present, while the sialylated glycans A2G1S1, A2G2S1, and A2G2S1 are correctly predicted to be experimentally absent in this altered terminal nucleotide glycoform engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

C.3.1.2 Strategy 2: Genetic Manipulation of Glycosyltransferases

Glyco-Mapper successfully modeled glycosyltransferase glycoform-engineering studies in addition to *GnT-I* expression by Goh (Goh et al. 2014) and *Fut8* knockout by Kanda (Kanda et al. 2007) including *GnT-I* knockout (Sealover et al. 2013), *Fut8* knockouts via alternative methods (Tsukahara et al. 2006; Malphettes et al. 2010), and *β4Galt* and *ST3Gal3* overexpression (Weikert et al. 1999). In contrast to Goh, Sealover (Sealover et al. 2013) investigated the effect of reduced *GnT-I* expression in a CHO cell line with the goal of decreasing the IgG mAb glycoform variety using a zinc-finger nuclease (ZFN) *GnT-I* knockout. Glyco-Mapper accurately modeled 37 of 40 glycans [3 of 4 present; 34 of 36 absent] for the wild type glycoform (Figure C.3). Glyco-Mapper predicted a different 37 of 40 glycans correctly [0 of 1 present, 37 of 39 absent] and data confirmed each of the 3 glycans that changed prediction classes when the *GnT-I* knockout was modeled (Figure C.4). The predicted glycosyltransferase knockout mAb glycoform was highly accurate and specific.

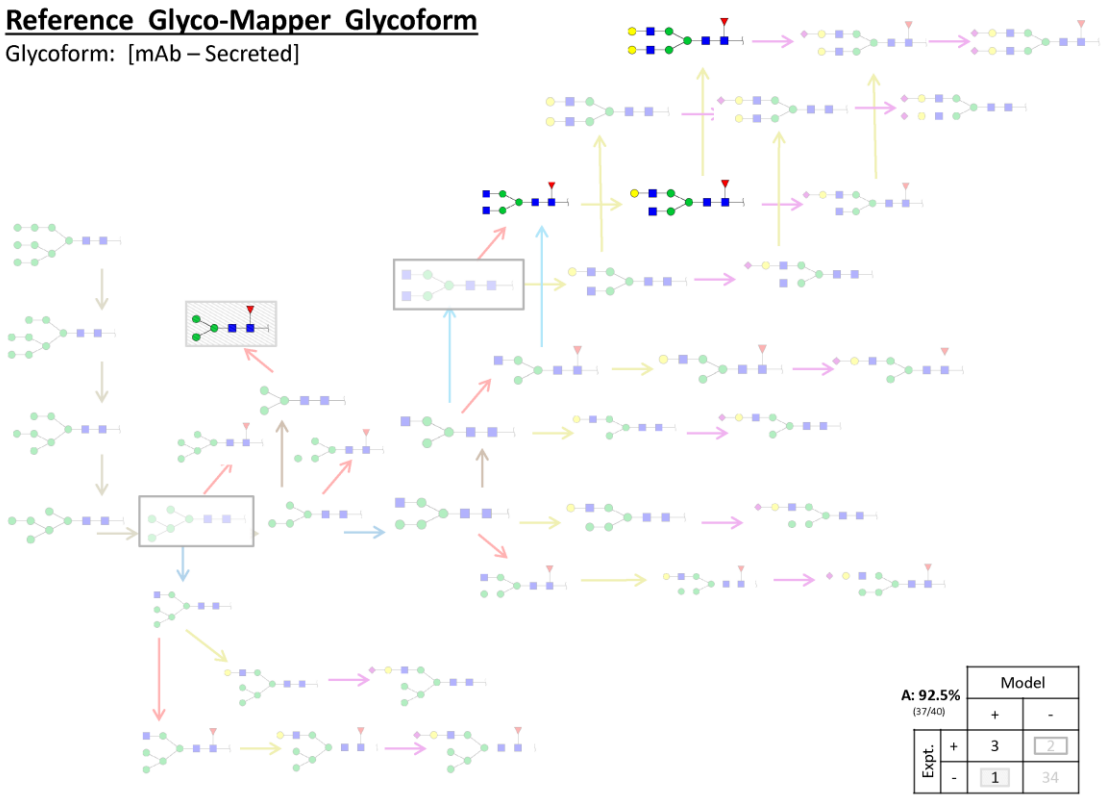


Figure C.3: The Glyco-Mapper prediction of the Sealover *et al.* reference glycoform. The fucosylated bi-antennary glycans FA2, FA2G1, and FA2G2 are correctly predicted to be experimentally present in this reference glycoform. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

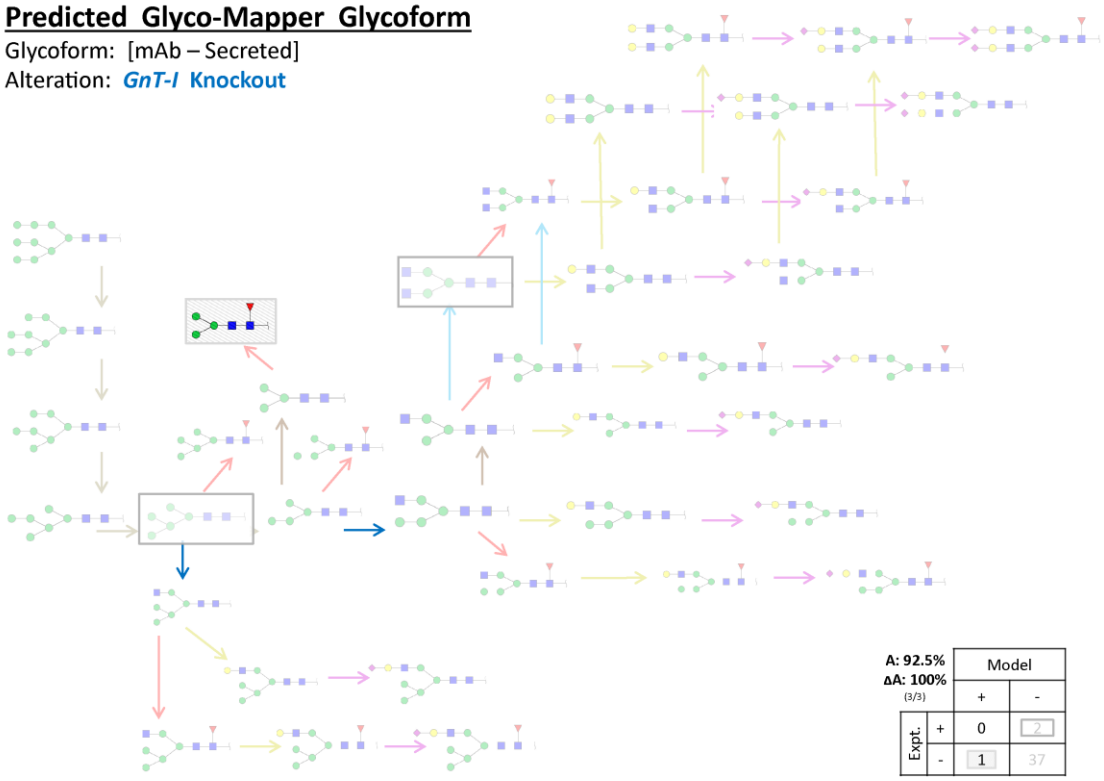


Figure C.4: The Glyco-Mapper prediction of the *GnT-I* knockout strategy based on the Sealover *et al.* reference glycoform (Figure C.3). The fucosylated bi-antennary glycans FA2, FA2G1, and FA2G2 are all correctly predicted to be experimentally absent in this glycoform-engineering strategy. The glycans M5 and A2 that were incorrectly predicted to be absent in the reference glycoform are still incorrectly predicted to be absent with the *GnT-I* knockout. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

Malphettes (Malphettes et al. 2010) investigated the effect of a ZFN *Fut8* knockout in a CHO cell line with the goal of creating afucosylated IgG1 mAbs. Glyco-Mapper accurately modeled 37 of 40 glycans [6 of 8 present; 31 of 32 absent] for the wild type glycoform (Figure C.5). Glyco-Mapper predicted 37 of 40 glycans correctly [6 of 8 present, 31 of 32 absent] and data confirmed 7 of the 8 glycans that changed

prediction classes when the *Fut8* knockout was modeled (Figure C.6). The predicted afucosylated mAb glycoform was highly accurate and specific.

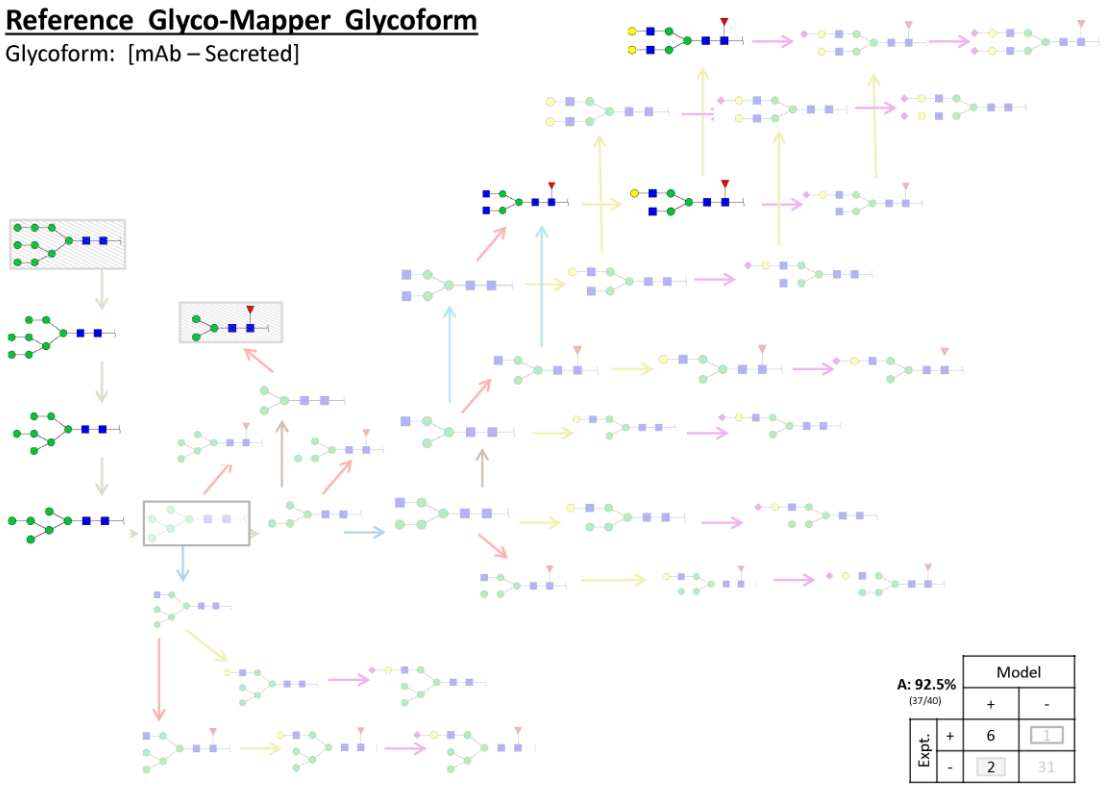


Figure C.5: The Glyco-Mapper prediction of the Malphettes *et al.* reference glycoform. Multiple high mannose (M8, M7, and M6) and bi-antennary (FA2, FA2G1, FA3G3S2, and FA2G2) glycans are all correctly predicted to be experimentally present in this reference glycoform. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

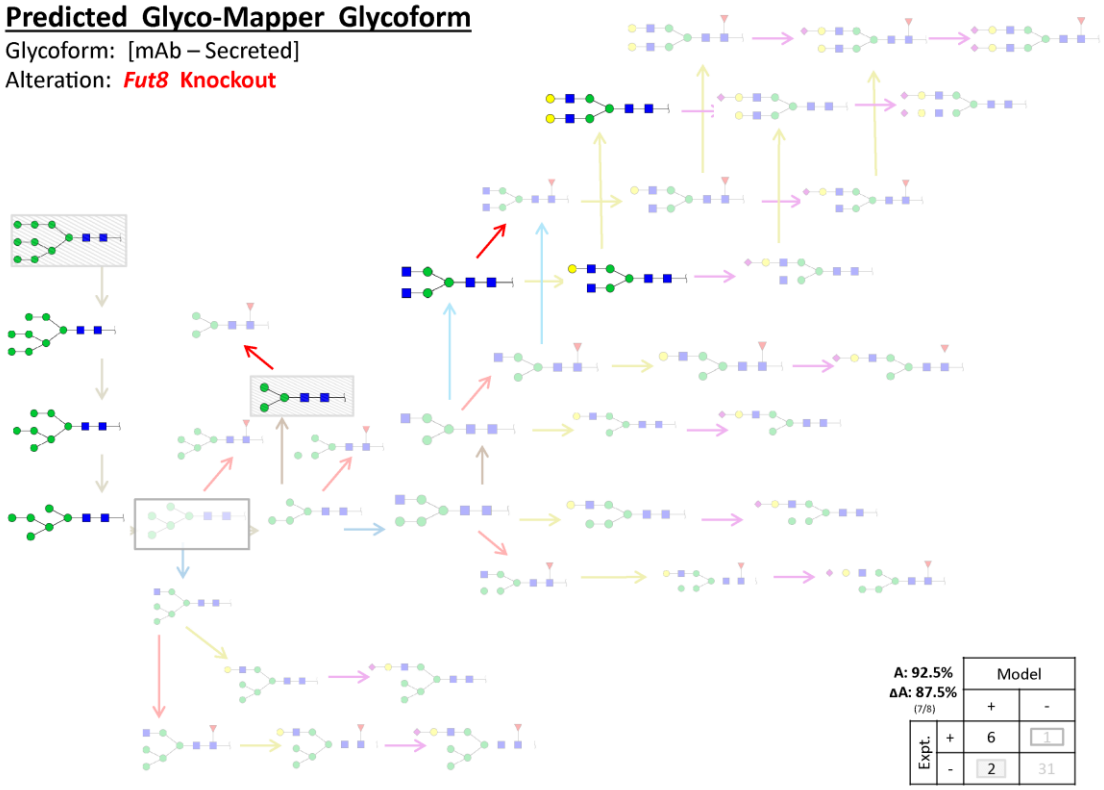


Figure C.6: The Glyco-Mapper prediction of the knockout of *Fut8* based on the Malphettes *et al.* reference glycoform (Figure C.5). The afucosylated bi-antennary glycans A2, A2G1, and A2G2 are correctly predicted to be experimentally present and the fucosylated bi-antennary glycans FA2, FA2G1, and FA2G2 are correctly predicted to be experimentally absent in this glycosyltransferase glycoform-engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

Tsukahara (Tsukahara et al. 2006) also investigated the effect of *Fut8* knockout by homologous recombination in a CHO cell line producing a mAb biotherapeutic. Glyco-Mapper accurately modeled 38 of 40 glycans [6 of 8 present; 32 of 32 absent] for the wild type glycoform (Figure C.7) and predicted 39 of 40 glycans correctly [3 of 4 present, 36 of 36 absent] and data confirmed each of the 4 glycans

that changed prediction classes when the *Fut8* knockout was modeled (Figure C.8). The predicted afucosylated mAb glycoform was highly accurate and specific.

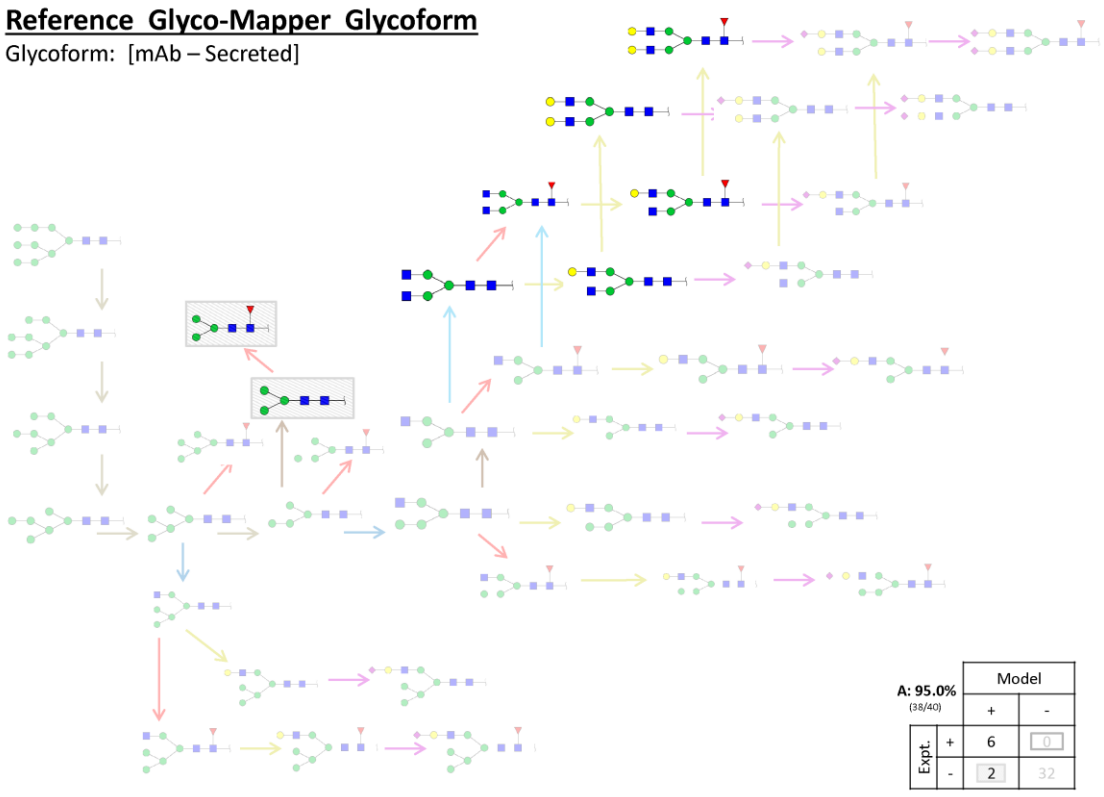


Figure C.7: The Glyco-Mapper prediction of the Tsukahara *et al.* reference glycoform. The asialylated, bi-antennary glycans A2, FA2, A2G1, FA2G1, A2G2, and FA2G2 are all correctly predicted to be experimentally present in this reference glycoform. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

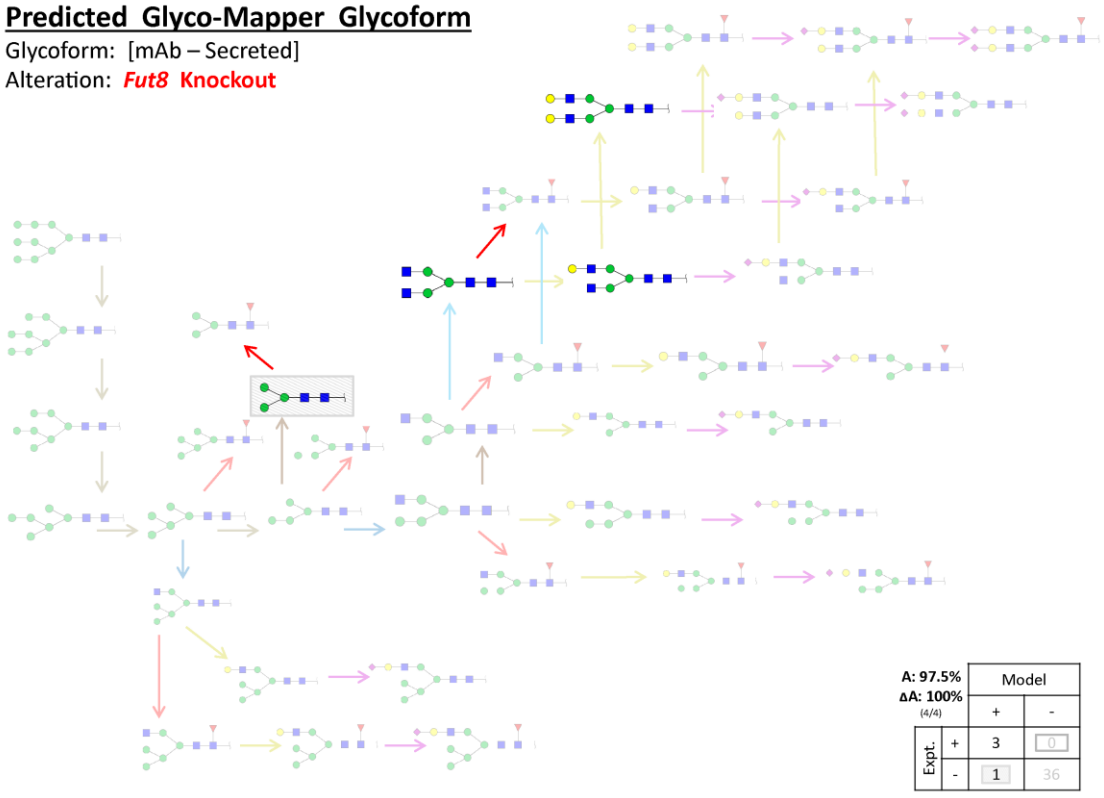


Figure C.8: The Glyco-Mapper predicted *Fut8* knockout glycoform is based on the Tsukahara *et al.* reference glycoform (Figure C.7). The afucosylated glycans A2, A2G1, and A2G2 are all correctly predicted to be experimentally present while the fucosylated glycans FA2, FA2G1, and FA2G2 are all correctly predicted to be experimentally absent in this glycoform-engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

Weikert (Weikert et al. 1999) investigated the effects of *β4Galt* and *ST3Gal3* expression in a CHO cell line with the goal of understanding the effects of gene overexpression on EPO galactosylation and sialylation. Glyco-Mapper accurately modeled 150 of 156 glycans [8 of 13 present; 142 of 143 absent] for the wild type glycoform (Figure C.9). When the *β4Galt* overexpression was modeled, Glyco-

Mapper predicted 149 of 156 glycans correctly [7 of 13 present, 142 of 143 absent] (Figure C.10); whereas, when the *ST3Gal3* overexpression was modeled, Glyco-Mapper predicted 148 of 156 glycans correctly [6 of 13 present, 142 of 143 absent] (Figure C.11). Glyco-Mapper accurately modeled 150 of 156 glycans [9 of 13 present; 141 of 143 absent] for a second, different wild type glycoform (Figure C.12). Glyco-Mapper again predicted 150 of 156 glycans correctly [9 of 13 present, 141 of 143 absent] when the *β4Galt* and *ST3Gal3* overexpression was modeled (Figure C.13). The predicted EPO glycoform resulting from the altered glycosyltransferase expression was highly accurate and specific, but largely unaltered.

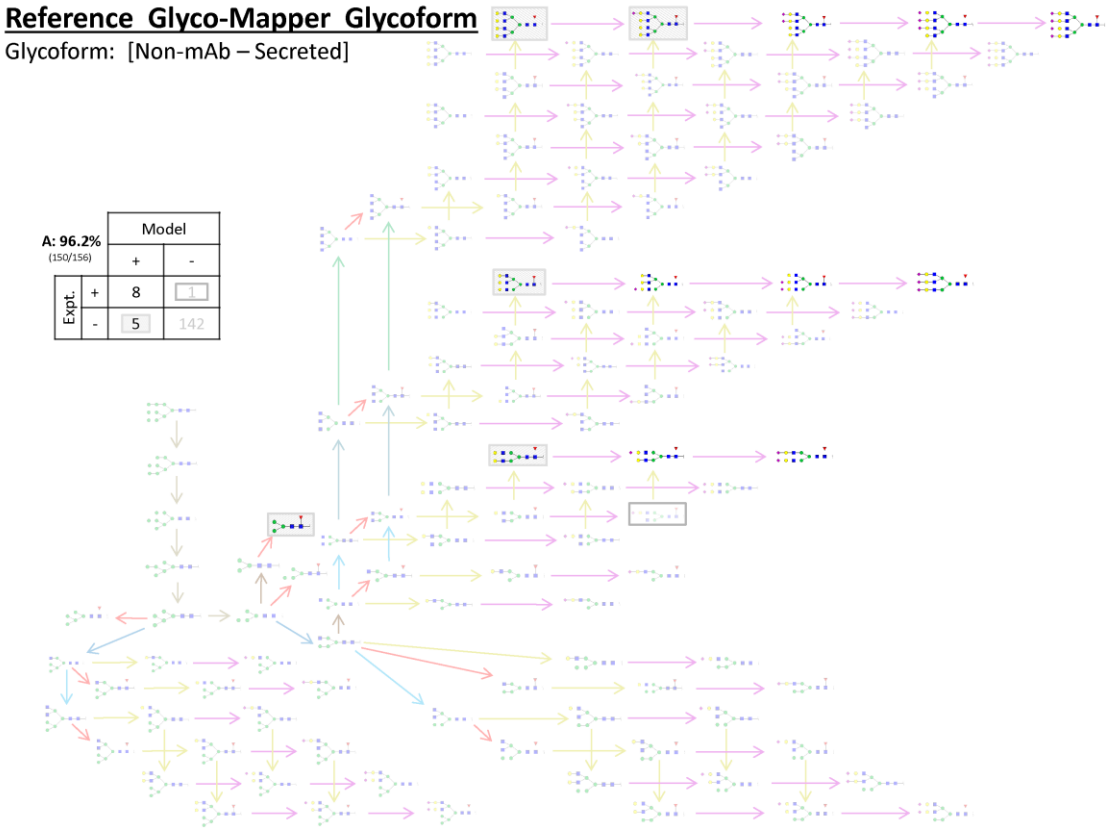


Figure C.9: The Glyco-Mapper prediction of the first Weikert *et al.* reference glycoform. The fucosylated, bi-antennary (FA2G2S1 and FA2G2S2), tri-antennary (FA3G3S1, FA3G3S2, and FA3G3S3), and tetra-antennary (FA4G4S2, FA4G4S3, and FA4G4S4) glycans are all correctly predicted to be experimentally present in this reference glycoform. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

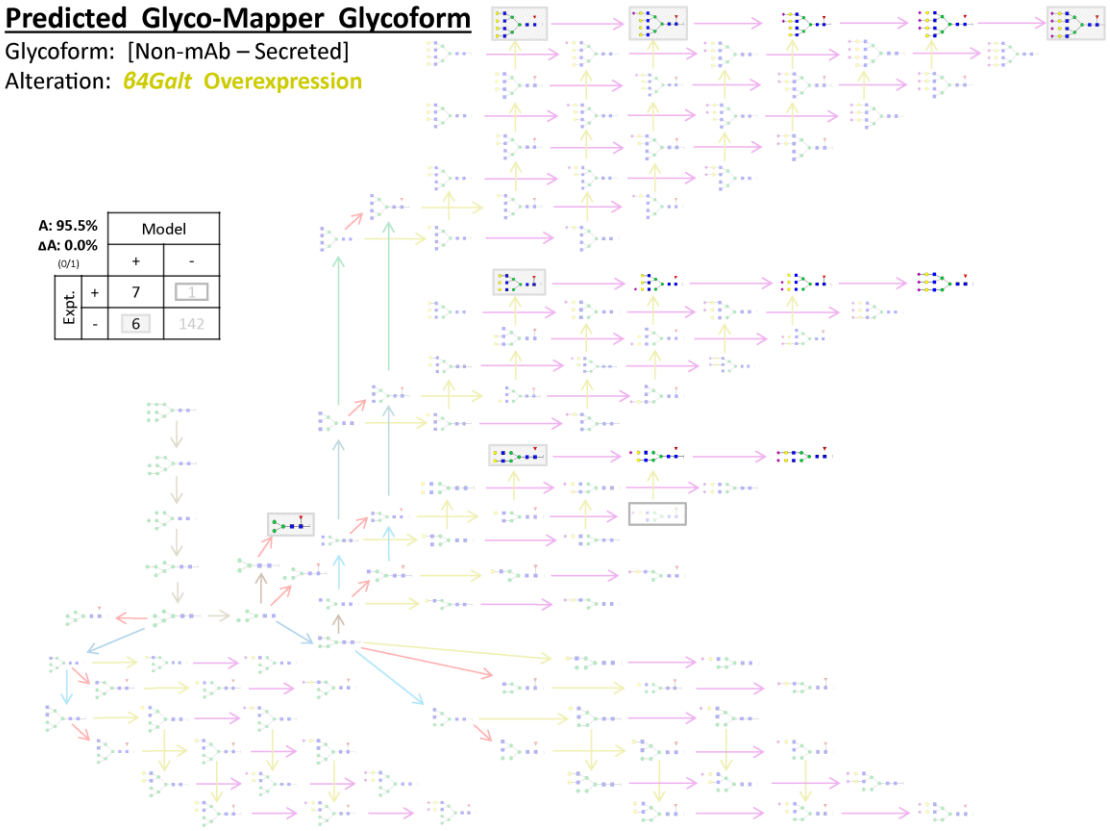


Figure C.10: The Glyco-Mapper predicted *β4Galt* overexpression glycoform is based on the initial Weikert *et al.* reference glycoform (Figure C.9). The fucosylated bi-antennary (FA2G2S1 and FA2G2S2), tri-antennary (FA3G3S1, FA3G3S2, and FA3G3S3), and tetra-antennary (FA4G4S2 and FA4G4S3) glycans are all correctly predicted to be experimentally present, but the glycan FA4G4S4 is now incorrectly predicted to be experimentally absent in this glycoform-engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

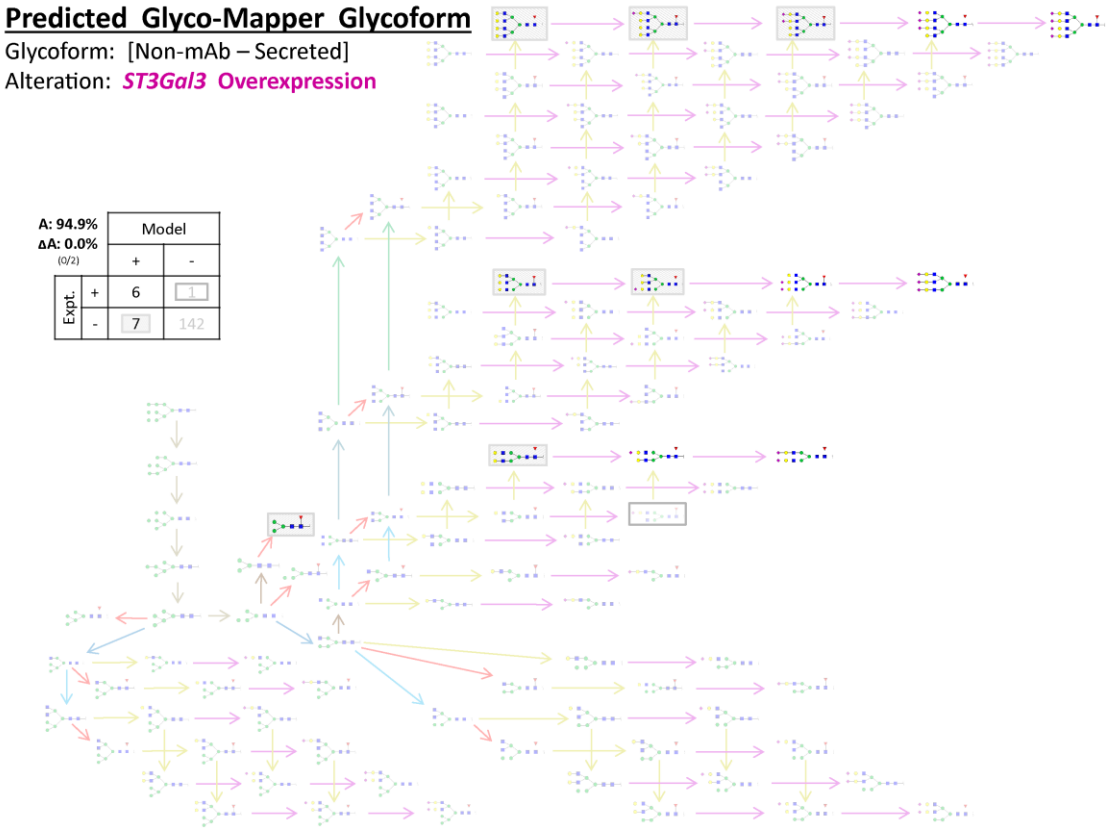


Figure C.11: The Glyco-Mapper prediction of the *ST3Gal3* overexpression glycoform based on the initial Weikert *et al.* reference glycoform (Figure C.9). The fucosylated and highly-sialylated bi-antennary (FA2G2S1 and FA2G2S2), tri-antennary (FA3G3S2 and FA3G3S3), and tetra-antennary (FA4G4S3 and FA4G4S4) glycans are all correctly predicted to be experimentally present in this overexpressed glycosyltransferase glycoform-engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

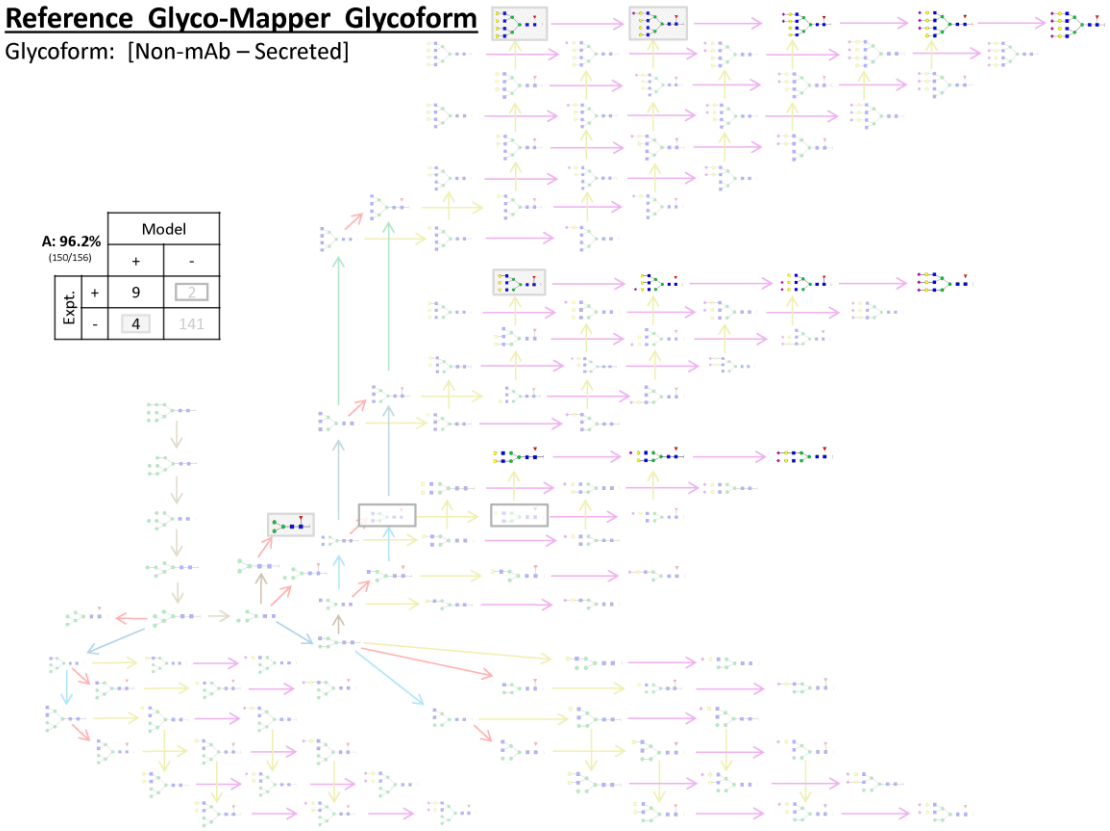


Figure C.12: The Glyco-Mapper prediction of the second Weikert *et al.* reference glycoform. The fucosylated, bi-antennary (FA2G2, FA2G2S1, and FA2G2S2), tri-antennary (FA3G3S1, FA3G3S2, and FA3G3S3), and tetra-antennary (FA4G4S2, FA4G4S3, and FA4G4S4) glycans are all correctly predicted to be experimentally present in this reference glycoform. The difference between this reference glycoform and the initial reference glycoform (Figure C.9), besides the glycoforms originating under different conditions, is the glycan FA2G2 is experimentally measured. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

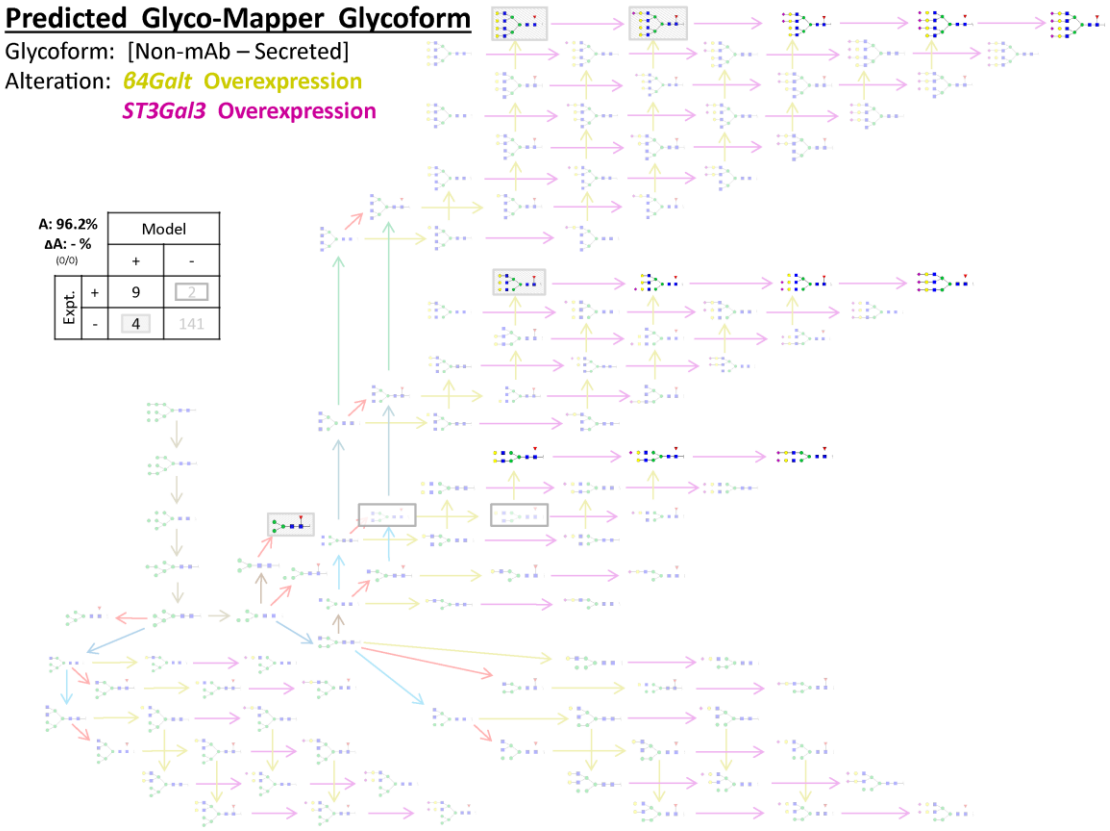


Figure C.13: The Glyco-Mapper prediction of the coupled overexpression of *β4Galt* and *ST3Gal3* based on the second Weikert *et al.* reference glycoform (Figure C.12). The fucosylated, bi-antennary (FA2G2, FA2G2S1, and FA2G2S2), tri-antennary (FA3G3S1, FA3G3S2, and FA3G3S3), and tetra-antennary (FA4G4S2, FA4G4S3, and FA4G4S4) glycans are all correctly predicted to be experimentally present in this multiple glycosyltransferase overexpression glycoform. There is not one difference between this glycoform prediction and the reference glycoform, which is a good outcome because the experimental glycoforms also did not contain any differences. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

C.3.1.3 Strategy 3: Genetic Manipulation of Glycosyltransferase and Metabolism Genes and Nutrient Feeding Modifications

The knockout of a native metabolism gene or the alteration of a media feed (nutrient composition) can result in a modified glycoform, as reported by Kanda (Kanda et al. 2007) and Imai-Nishiya (Imai-Nishiya et al. 2007), both of which Glyco-Mapper successfully modeled. Using a mAb (IgG1)-producing CHO cell line, Imai-Nishiya et al. knocked out *GMDS* and *Fut8* simultaneously, thereby affecting the fucosylation and antibody-dependent cellular cytotoxicity through both the glycosylation and metabolism processes. Glyco-Mapper modeled the wild type glycoform (Figure C.14) with 39 of 40 correct glycans [3 of 4 present, 36 of 36 absent]. When the *GMDS* and *Fut8* knockout was modeled (Figure C.15), Glyco-Mapper accurately predicted 39 of 40 glycans correctly [3 of 4 present; 36 of 36 absent] and data confirmed 7 of the 8 glycans that changed prediction classes. The predicted mAb glycoform resulting from the altered metabolic and glycosyltransferase gene expression was highly accurate, sensitive, and specific.

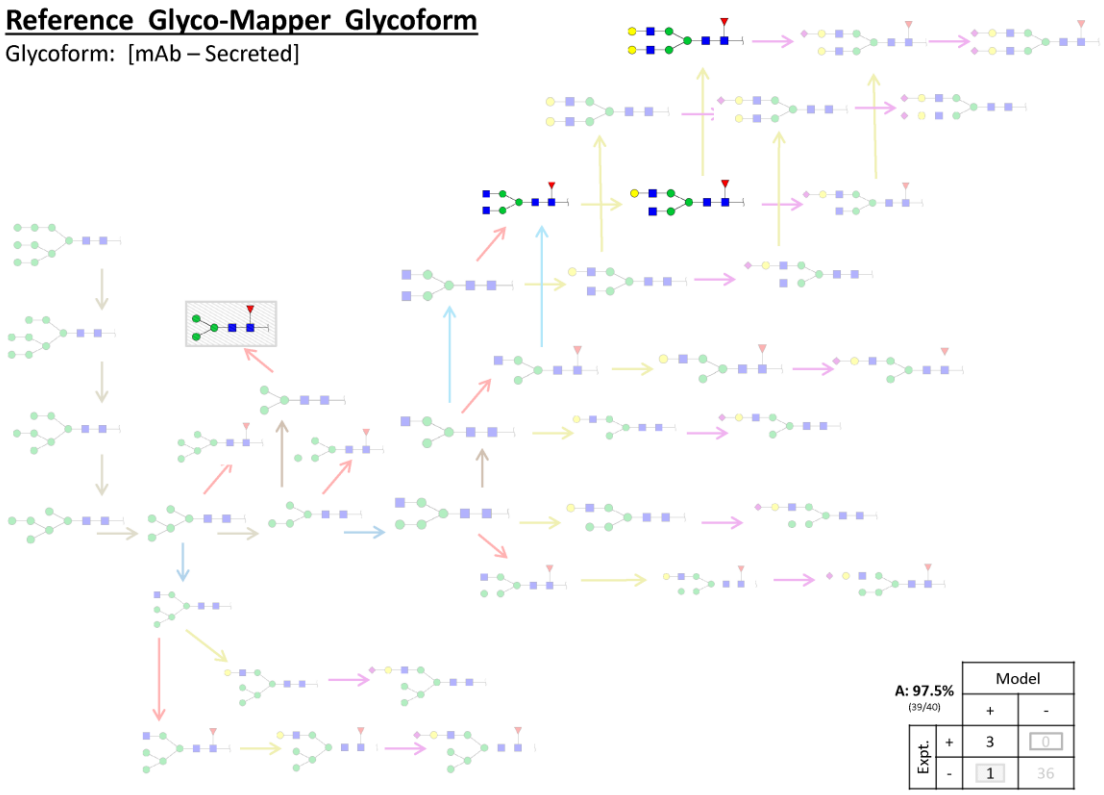


Figure C.14: The Glyco-Mapper prediction of the Imai-Nishiya *et al.* reference glycoform. The fucosylated bi-antennary glycans FA2, FA2G1, and FA2G2 are correctly predicted to be experimentally present in this reference glycoform. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

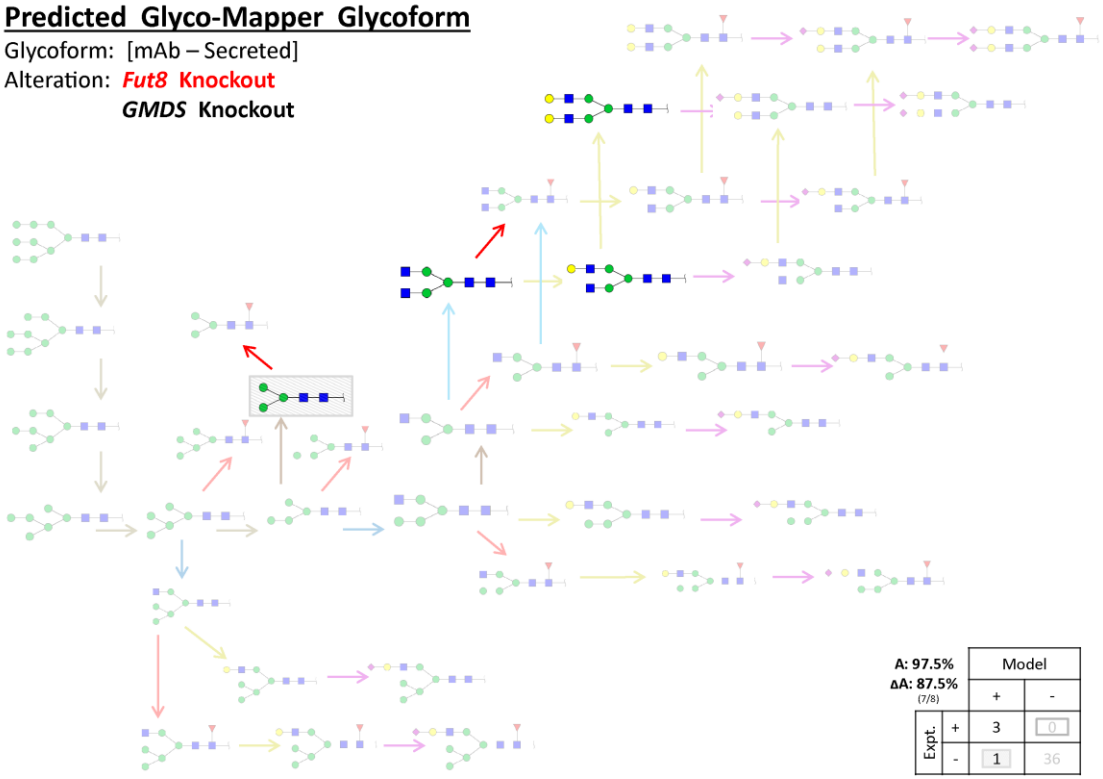


Figure C.15: The Glyco-Mapper predicted *Fut8* and *GMDS* knockout glycoform is based on the Imai-Nishiya *et al.* reference glycoform (Figure C.14). The afucosylated bi-antennary glycans A2, A2G1, and A2G2 are correctly predicted to be experimentally present and the fucosylated bi-antennary glycans FA2, FA2G1, and FA2G2 are correctly predicted to be experimentally absent in this glycosyltransferase and metabolism-based glycoform-engineering strategy. The N-glycan gene, nucleotide, and glycan legends in Figure C.1 are not pictured but still apply.

C.3.2 Prediction of Uncommon Published Glycoforms

The Ggta enzyme is responsible for creating the Gal-α-1,3-Gal (α-Gal) linkage (Taniguchi et al. 2002). Terminal α-Gal nucleotides can react with the α-Gal antibodies commonly produced in humans (Macher and Galili 2008). The α-Gal antigen is rarely measured in CHO cell proteins, yet a *Ggta* ortholog has been reported in the CHO-K1 genome (Bosques et al. 2010) and the α-Gal antigen has been previously reported in the CHO-produced antibody Erbitux (centuximab) (Chung et al. 2008). The Glyco-Mapper prediction of a glycoform containing an α-Gal glycan represents the α-Gal with G* and this linkage competes with the sialyltransferase enzymes for terminal Gal nucleotides.

Another family of N-glycosylation genes that is annotated in the CHO genome (Xu et al. 2011) but rarely affect biopharmaceutical glycoforms is the *β4Galnt* gene family. The β4Galnt3 and β4Galnt4 enzymes are responsible for creating the GalNAc-β-1,4-GlcNAc linkage (Taniguchi et al. 2002), which is a terminal cell surface N-glycan linkage (Ikehara et al. 2006). GalNAc is represented by "L" in the Glyco-Mapper tool and the β4Galnt3 and β4Galnt4 enzymes compete with the galactosyltransferase enzymes for the terminal GlcNAc nucleotides.

C.4 Discussion

C.4.1 Glyco-Mapper Successes

The biopharmaceutical protein structure and the CHO cell line producing the biopharmaceutical both affect the reference glycoform and are two causes for the glycoform variety reported in literature. Glyco-Mapper accounted for these parameters and successfully modeled various reference glycoforms, enabling accurate cell-engineering glycoform predictions regardless of the reference glycoform's number of reported glycans, number of subclasses represented by these glycans, or the maximum number of reactions between these glycans. The reference glycoforms ranged in the number of experimentally confirmed glycans from 3 glycans (Figure 4.1) to 20 glycans (Figure 5.1), demonstrating Glyco-Mapper's ability to model glycoforms both large and small. The reference glycoforms also ranged in the number of correctly modeled glycan subclasses from only one bi-antennary subclass (Figure 4.5) to five subclasses (Figure 4.9), including high- and low-mannose as well as bi-, tri-, and tetra-antennary subclasses. This range in the number of modeled glycan subclasses highlights the span of glycan structures that were able to be simultaneously represented. The range in the maximum number of enzymatic reactions between two measurable glycans in the same glycoform ranged from 2 reactions catalyzed by 1 enzyme (Figure 4.1) to 15 reactions catalyzed by 8 different enzymes (Figure 4.9). The range of modeled glycans from only 1 enzyme's reactants and products to a varied multitude of enzyme products established the Glyco-Mapper's significant

modeling versatility. This ability to adapt the reference glycoform to closely resemble the unique experimental reference glycoform enables accurate cell line- and biopharmaceutical-specific predictions to be made, greatly increasing the tool's specificity and applicability.

Glyco-Mapper successfully predicted many glycoforms in part because of the customized reference glycoforms. The predicted glycoforms were characterized by their quantized glycoform patterns, the vast range of gene alteration glycan results achievable, and multiple predicted glycoforms depending on input variation. The prediction of quantized glycoform patterns consistent with experimental data was demonstrated with a distinct non-mAb pattern in Figure 4.10, contrasting a smaller, but also distinct mAb pattern in Figure C.6. The variety of gene alterations that affect the glycan results ranged from significant composition changes due to *GnT* gene alterations (Figures 4.2 and C.4) to subtle, yet biopharmaceutically relevant changes due to *Fut8* or *STGal* gene alterations (Figures 4.7 and C.2). The prediction outcomes vary depending on the discretized parameter value, as demonstrated by the lack of any predicted glycan changes in Figure C.13 to the 15 predicted glycan changes in Figure 4.4, which is indicative of an accurate representation of variable sensitivity. These benefits highlight the wide variety of glycoforms that can be applicably modeled by Glyco-Mapper.

C.4.2 Current Glyco-Mapper Challenges

As shown in Table 4.4, 24 glycans from the five figures listed were incorrectly predicted, totaling 4% of the predicted glycans within those figures, and this is representative of all 17 predicted glycoforms with an incorrect glycan prediction rate

of 4% as well. The two Glyco-Mapper prediction errors were glycans that were incorrectly predicted to be present within the glycoform and glycans that were incorrectly predicted to be absent from the glycoform. The glycan patterns portrayed using Glyco-Mapper were highly accurate, but the erroneous predictions may be indicative of trends that might not otherwise be noticed. A few of these trends are thoroughly examined here.

C.4.2.1 Incorrectly Predicted Glycans Indicative of Potential Network Uncertainties

Incorrect Glyco-Mapper predictions may identify reactions that are not accurately understood due to unaccounted for inhibiting factors. The example in Chapter 4 involved glycan A2G2S2 in Figure 4.2 and is indicative of a potential relationship between fucosylation and sialylation. Another example is from the Maszczak-Seneczko glycoforms (Figures 4.9-4.12) dealing with a potential relationship between fucosylation and galactosylation or antennarity. In each of the figures, there are at least two fucosylated/afucosylated glycan pairs of different antennarity, but only the fucosylated glycan of at least one pair is experimentally present in three of the glycoforms: twice in Figure 4.9, twice in Figure 4.10, and once in Figure 4.12. The Maszczak-Seneczko cell-engineered alterations involved the knockout of *β4Galt* in both Figures 4.10 and 4.12 and as galactosyltransferases and sialyltransferases are both located in the trans-Golgi (Hassinen and Kellokumpu 2014), there may be also be a correlation between galactosylation and fucosylation as theorized between sialylation and fucosylation. Regardless of whether this pattern indicates a definitive mechanism or simply a trend between fucosylation and

galactosylation or antennarity, this pattern may not have been detected without the Glyco-Mapper's consistent incorrect prediction of these glycans.

C.4.2.2 Incorrectly Predicted Glycans 1-2 Modifications "Off-Target"

Many glycans the Glyco-Mapper incorrectly predicted as present are only one or two active enzyme modifications removed from a glycan correctly predicted as present (Table 4.4). Amongst the Maszczak-Seneczko et al. reference and predicted glycoforms (Figures 4.9-4.12), there were a total of 15 incorrectly predicted glycans and all 15 glycans were only one modification away from a glycan correctly predicted to be present. Upon examination of this group of incorrect predictions, these glycans are likely the result of the non-template driven enzymatic reaction network. The Glyco-Mapper's high overall modeling accuracy is further strengthened by this trend because many incorrect glycan predictions are one to two modifications from a correctly predicted glycan, demonstrating a small degree of inaccuracy between highly similar glycans rather than a significant degree of inaccuracy between vastly different glycans.

C.4.2.3 Incorrectly Predicted Intermediate Glycans with Reactant and Product Glycans

A subset of the incorrectly predicted glycans that are 1-2 modifications removed from a glycan correctly predicted as present are one modification removed from that glycan's respective reactant and product glycans, both of which are correctly predicted to be present. Three different experimental glycoforms contain incorrectly predicted glycans that classify as an incorrectly predicted "intermediate" glycan and

each glycan is bi-antennary and has at least one terminal N-acetylglucosamine (GlcNAc) nucleotide. Multiple other bi-antennary glycans with terminal GlcNAc nucleotides are correctly predicted to be present within each glycan's respective glycoform (2 for A2G1S1 (Figure C.1), 3 for FA2 (Figure 4.5), and 5 for A2G1S1 (Figure 5.1)). A pattern does not appear between these few examples regarding why certain bi-antennary, GlcNAc-terminal glycans are measured in some circumstances and not others. One possible theory can be developed from the glycan A2G1S1, which I detected within the reference SEAP glycoform (Figure 5.1), yet the quantified measurement was not statistically significant and was excluded from analysis. This highlights the possibility that these incorrectly predicted glycans may actually be present in the experimental glycoforms, but in too low an amount to be statistically significant for analysis due to the presence of the other similar glycans only one modification away.

C.4.2.4 Incorrectly Predicted Mannose Glycans

Mannose glycans are present within biopharmaceutical glycoforms and while commonly accounting for a small portion of the glycoform compared to the complex glycans, mannose glycans also need to be correctly predicted and modeled by Glyco-Mapper. However, mannose glycans are erroneously predicted by the Glyco-Mapper. The mannose glycan M5 is incorrectly predicted to be absent for each of the reference and predicted glycoforms listed: Goh (Figures 4.3-4.4), Maszczak-Seneczko (Figures 4.9-4.12), Sealover (Figures C.3-C.4), Malphettes (Figures C.5-C.6), and the novel SEAP *GnT-II* knockdown (Figures 4.13-4.14). Within the Goh and Sealover *GnT-I* knockout glycoforms, the complex glycan production pathway is nonfunctional,

allowing Man-II to create the low-mannose M3 glycan. Glyco-Mapper predictions accurately represent this outcome but predict the complete consumption of M5. Within the modeled and predicted Maszczak-Seneczko glycoforms, GnT-I and Man-II are both enzymatically active and the M5 glycan is again predicted to be completely consumed as an intermediate. The M3 glycan is frequently incorrectly predicted by the Glyco-Mapper in the Naso (Figures C.1-C.2), Tsukahara (Figures C.7-C.8), Imai-Nishiya (Figure C.15), and Kanda (Figures 4.5-4.7) glycoforms. Man-II is required to produce complex glycans, but there are many substrates Man-II can act upon, including M5, M4, and infrequent hybrid glycans. This results in the correct prediction of M3 for many cases, but also incorrect predictions for others. A better understanding of these pathways would enable more accurate modeling and predictions.

C.4.3 Uncommon Glycoforms

The α-Gal linkage produced by the Ggta enzyme and the GalNAc-β-1,4-GlcNAc linkage produced by the β4Galnt3 and β4Galnt4 enzymes are both uncommon N-glycosylation linkages within biopharmaceuticals. The measured biopharmaceutical production of the α-Gal linkage and sequenced *Ggta* ortholog warranted the inclusion of this gene, enzyme, and linkage into the Glyco-Mapper. This linkage is undesirable and uncommon outside of CHO-produced IgE biopharmaceuticals (Chung et al. 2008), and while it will likely be of little interest to most Glyco-Mapper users, the *Ggta* ortholog and α-Gal linkage were included in the Glyco-Mapper's glycosylation database. The GalNAc-β-1,4-GlcNAc linkage is not measured within biopharmaceuticals and will also be of minimal interest to the majority of users, yet the annotated *β4Galnt3* and *β4Galnt4* genes are N-glycosylation

genes and are also represented within the CHO Glyco-Mapper tool's glycosylation database.

REFERENCES

Bosques CJ, Collins BE, Meador JW, Sarvaiya H, Murphy JL, DelloRusso G, Bulik DA, Hsu IH, Washburn N, Sipsey SF, Myette JR, Raman R, Shriver Z, Sasisekharan R, Venkataraman G. (2010) Chinese hamster ovary cells can produce galactose-α-1,3-galactose antigens on proteins. Nat Biotechnol. 28:1153-1156.

Chung CH, Mirakhur B, Chan E, Le QT, Berlin J, Morse M, Murphy BA, Satinover SM, Hosen J, Mauro D, Slebos RJ, Zhou Q, Gold D, Hatley T, Hicklin DJ, Platts-Mills TAE. (2008) Cetuximab-induced anaphylaxis and IgE specific for galactose-alpha-1,3-galactose. N Engl J Med. 358:1109-1117.

Goh JSY, Liu Y, Chan KF, Wan C, Teo G, Zhang P, Zhang Y, Song Z. (2014) Producing recombinant therapeutic glycoproteins with enhanced sialylation using CHO-gmt4 glycosylation mutants. Bioengineered. 5:1–5.

Hassinen A, Kellokumpu S. (2014) Organizational interplay of Golgi N-glycosyltransferases involves organelle microenvironment-dependent transitions between enzyme homo- and heteromers. J Biol Chem. 289:26937-26948.

Ikehara Y, Sato T, Niwa T, Nakamura S, Gotoh M, Ikehara SK, Kiyohara K, Aoki C, Iwai T, Nakanishi H, Hirabayashi J, Tatematsu M, Narimatsu H. (2006) Apical Golgi localization of N,N'-diacetyllactosediamine synthase, beta4GalNAc-T3, is responsible for LacdiNAc expression on gastric mucosa. Glycobiology. 16:777-785.

Imai-Nishiya H, Mori K, Inoue M, Wakitani M, Iida S, Shitara K, Satoh M. (2007) Double knockdown of a1,6-fucosyltransferase (FUT8) and GDP-mannose 4,6-dehydratase (GMD) in antibody-producing cells: a new strategy for generating fully non-fucosylated therapeutic antibodies with enhanced ADCC. BMC Biotechnol. 7:84-96.

Kanda Y, Imai-Nishiya H, Kuni-Kamochi R, Mori K, Inoue M, Kitajima-Miyama K, Okazaki A, Iida S, Shitara K, Satoh M. (2007) Establishment of a GDP-mannose 4,6-dehydratase (GMD) knockout host cell line: A new strategy for generating completely non-fucosylated recombinant therapeutics. J Biotechnol. 130:300–310.

Macher BA, Galili U. (2008) The Gal alpha 1,3Gal beta 1,4GlcNAc-R(alpha-Gal) epitope: a carbohydrate of unique evolution and clinical relevance. Biochim Biophys Acta. 1780:75-88.

Malphettes L, Freyvert Y, Chang J, Liu PQ, Chan E, Miller JC, Zhou Z, Nguyen T, Tsai C, Snowden AW, Collingwood TN, Gregory PD, Cost GJ. (2010) Highly Efficient Deletion of FUT8 in CHO Cell Lines Using Zinc-Finger Nucleases Yields Cells That Produce Completely Nonfucosylated Antibodies. Biotechnol Bioeng. 130:300–310.

Maszczak-Seneczko D, Sosicka P, Olczak T, Jakimowicz P, Majkowski M, Olczak M. (2013) UDP-N-acetylglucosamine Transporter (SLC35A3) Regulates Biosynthesis of Highly Branched N-glycans and Keratan Sulfate. J Biol Chem. 288:21850–21860.

Naso MF, Tam SH, Scallon BJ, Raju TS. (2010) Engineering host cell lines to reduce terminal sialylation of secreted antibodies. mAbs. 2:519-527.

Onitsuka M, Kim WD, Ozaki H, Kawaguchi A, Honda K, Kajiura H, Fujiyama K, Asano R, Kumagai I, Ohtake H, Omasa T. (2012) Enhancement of sialylation on humanized IgG-like bispecific antibody by overexpression of α-2,6-sialyltransferase derived from Chinese hamster ovary cells. Biotechnol Prod Proc Eng. 94:69–80.

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

Tsukahara M, Aoki A, Kozono K, Maseki Y, Fukuda Y, Yoshida H, Kobayashi K, Kakitani M, Tomizuka K, Tsumura H. (2006) Targeted disruption of α-1,6-fucosyltransferase (FUT8) gene by homologous recombination in Chinese hamster ovary (CHO) cells. Animal Cell Technol. 14:175-183.

Weikert S, Papac D, Briggs J, Cowfer D, Tom S, Gawlitzek M, Lofgren J, Mehta S, Chisholm V, Modi N, Eppler S, Carroll K, Chamow S, Peers D, Berman P, Krummen L. (1999) Engineering Chinese hamster ovary cells to maximize sialic acid content of recombinant glycoproteins. Nat Biotechnol. 17:1116-1121.

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

Yin B, Gao Y, Chung CY, Yang S, Blake E, Stuczynski MC, Tang J, Kildegaard HF, Anderson MR, Zhang H, Betenbaugh MJ. (2015) Glycoengineering of Chinese hamster ovary cells for enhanced erythropoietin N-glycan branching and sialylation. Biotechnol Bioeng. 112:2343-2351.

Appendix D

REPRINT PERMISSIONS

D.1 Reprint Permissions for Chapter 3

Publication: B. G. Kremkow, J. Y. Baik, M. L. MacDonald, K. H. Lee. CHOgenome.org 2.0: Genome resources and website updates. *Biotechnology Journal.* 2015;10:931-938.

Permission:

This is a License Agreement between Benjamin G Kremkow and John Wiley and Sons Inc provided by Copyright Clearance Center.

Table D.1: License Agreement for Chapter 3.

License number	3950831280271
License date	09/16/16
License content publisher	John Wiley and Sons Inc
Licensed content publication	Biotechnology Journal
Licensed content title	CHOgenome.org 2.0: Genome resources and website updates
Licensed copyright line	© 2015 Wiley Periodicals, Inc.
Licensed content author	Benjamin G. Kremkow, Jong Youn Baik, Madolyn L. MacDonald, Kelvin H. Lee
Licensed content date	07/07/15
Licensed content volume number	10
Licensed content issue number	7
Start page	931
End page	938
Requestor type	Other Published Product - "Republish in a thesis/dissertation"
Requestor type	Author of requested content
Portion	Full chapter/article
Title of thesis/dissertation	CHO-PRODUCED BIOPHARMACEUTICAL GLYCOFORM PREDICTIONS THROUGH DISCRETIZED REACTION NETWORK MODELING

D.2 Reprint Permissions for Appendix A

Publication: B. Kremkow, K. H. Lee. Sequencing technologies for animal cell culture research. Biotechnology Letters. 2015;37:55-65.

Permission:

This is a License Agreement between Benjamin G Kremkow and SPRINGER-VERLAG DORDRECHT provided by Copyright Clearance Center.

Table D.2: License Agreement for Appendix A.

License number	3947720631942
License date	09/14/16
License content publisher	SPRINGER-VERLAG DORDRECHT
Licensed content publication	Biotechnology Letters
Licensed content title	Sequencing technologies for animal cell culture research
Licensed copyright line	© 2015 Springer Publishing Co.
Licensed content author	Benjamin G. Kremkow, Kelvin H. Lee
Licensed content date	01/05/15
Licensed content volume number	37
Licensed content issue number	1
Start page	55
End page	65
Requestor type	Republish or display content - "Thesis/Dissertation"
Requestor type	Author of requested content
Portion	Full article
Title of thesis/dissertation	CHO-PRODUCED BIOPHARMACEUTICAL GLYCOFORM PREDICTIONS THROUGH DISCRETIZED REACTION NETWORK MODELING

Appendix E

GENE LISTS FROM CHAPTER 4

E.1 Preface

As described in Chapter 4, 59 glycosylation genes and 92 metabolism genes compose the N-glycosylation reaction network within CHO cells and are accounted for in the Glyco-Mapper. The glycosylation genes are given in Table E.1 and the metabolism genes are givent in Table E.2. The metabolic pathways represented by these genes are pictured in Figure E.1.

Table E.1: The Glyco-Mapper glycosylation gene database. The relevant glycosylation gene symbols, IDs, and names within the Glyco-Mapper database.

Symbol	**CHO-K1 RefSeq (RS) ID**	**CHO-K1 RS Name**
DPAGT1	100689054	dolichyl-phosphate (UDP-N-acetylglucosamine) N-acetylglucosaminephosphotransferase 1 (GlcNAc-1-P transferase)
ALG13	100754023	ALG13, UDP-N-acetylglucosaminyltransferase subunit
ALG14	100773644	ALG14, UDP-N-acetylglucosaminyltransferase subunit
ALG1	100773731	ALG1, chitobiosyldiphosphodolichol beta-mannosyltransferase
ALG02	100768412	ALG2, alpha-1,3/1,6-mannosyltransferase
ALG3	100772003	ALG3, alpha-1,3- mannosyltransferase
ALG09	100755062	ALG9, alpha-1,2-mannosyltransferase
ALG11	100771009	ALG11, alpha-1,2-mannosyltransferase
ALG12	100770096	ALG12, alpha-1,6-mannosyltransferase
ALG6	100753783	ALG6, alpha-1,3-glucosyltransferase
ALG8	100766150	ALG8, alpha-1,3-glucosyltransferase
ALG10	-	-
ALG10B	-	-
ALG5	100769679	dolichyl-phosphate beta-glucosyltransferase
DPM1	100689420	dolichyl-phosphate mannosyltransferase polypeptide 1, catalytic subunit
DPM3	100689451	dolichyl-phosphate mannosyltransferase polypeptide 3
DDOST	100755259	dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit (non-catalytic)
RPN1	100762811	ribophorin I
RPN2	103158732	ribophorin II
DAD1	100767387	defender against cell death 1
MOGS	100689098	mannosyl-oligosaccharide glucosidase, transcript variant X1
GANAB	100752162	glucosidase, alpha; neutral AB, transcript variant X1
MAN1A1	100767553	mannosidase, alpha, class 1A, member 1
MAN1A2	100770527	mannosidase, alpha, class 1A, member 2
MAN1B1	100756512	endoplasmic reticulum mannosyl-oligosaccharide 1,2-alpha-mannosidase
MAN1C1	100761494	mannosidase, alpha, class 1C, member 1
MAN2A1	100757006	mannosidase, alpha, class 2A, member 1
MAN2A2	100764014	mannosidase, alpha, class 2A, member 2
MAN2B1	100766505	mannosidase, alpha, class 2B, member 1

Table E.1 continued.

Symbol	CHO-K1 RefSeq (RS) ID	CHO-K1 RS Name
MANBA	103158768	mannosidase, beta A, lysosomal
MGAT1	100682529	mannosyl (alpha-1,3-)-glycoprotein beta-1,2-N-acetylglucosaminyltransferase
MGAT2	100753385	mannosyl (alpha-1,6-)-glycoprotein beta-1,2-N-acetylglucosaminyltransferase
MGAT4A	100766200	mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isozyme A
MGAT4B	100768637	mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isozyme B
MGAT5	100760162	mannosyl (alpha-1,6-)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase
MGAT5B	100771275	mannosyl (alpha-1,6-)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase, isozyme B
MGAT3	100689076	mannosyl (beta-1,4-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase
FUT8	100751648	fucosyltransferase 8 (alpha (1,6) fucosyltransferase)
FUCA1	100752441	fucosidase, alpha-L- 1, tissue
β4GALT1	100689430	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 1
β4GALT2	100689434	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 2
β4GALT3	100689346	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 3
ST3Gal3	100689187	ST3 beta-galactoside alpha-2,3-sialyltransferase 3
ST6Gal1	100689389	ST6 beta-galactosamide alpha-2,6-sialyltranferase 1
ST6Gal2	100763756	ST6 beta-galactosamide alpha-2,6-sialyltranferase 2
Neu1	100689373	sialidase 1 (lysosomal sialidase)
Neu2	100689301	sialidase 2 (cytosolic sialidase)
Neu3	100689090	sialidase 3 (membrane sialidase)
Neu4	100774175	sialidase 4
HEXA	100750552	hexosaminidase A (alpha polypeptide)
HEXB	100756951	beta-hexosaminidase subunit beta-like
AGA	100754017	aspartylglucosaminidase
β4GALNT3	100756528	beta-1,4-N-acetyl-galactosaminyl transferase 3
β4GALNT4	100758404	beta-1,4-N-acetyl-galactosaminyl transferase 4
GLB1	100767446	galactosidase, beta 1 (galactosidase, beta 1-like)
CHST8	-	

Table E.1 continued.

Symbol	CHO-K1 RefSeq (RS) ID	CHO-K1 RS Name
CHST9	100770625	carbohydrate (N-acetylgalactosamine 4-0) sulfotransferase 9
STS	-	
Ggta1	100754535	N-acetyllactosaminide alpha-1,3-galactosyltransferase

Table E.2: The glycosylation-relevant metabolism and nucleotide sugar transporter gene database. Contains the relevant metabolism gene symbols, names, and reaction numbers associated with Figure E.1.

Abbr.	**Enzyme**	**Rxn**
HK	Hexokinase, (glucokinase)	1
GPI	Glucose Phosphate Isomerase, (phosphoglucoisomerase)	2
PFK	Phosphofructokinase	3
FBP	Fructose-1,6-Bisphosphatase	4
ALDO	Fructose Bisphosphate Aldolase	5
TPI	Triose phosphate isomerase (TIM)	6
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	7
PGK	Phosphoglycerate kinase	8
PGAM	Phosphoglycerate mutase	9
ENO	Enolase	10
PK	Pyruvate kinase	11
-	[Pyruvate transporter]	12
LDH	Lactate dehydrogenase	13
PC	Pyruvate carboxylase	14
PCK	Phosphoenolpyruvate carboxykinase	15
PDH	Pyruvate dehydrogenase	16
CS	Citrate synthase	17
ACO	Aconitase (Aconitate hydratase I)	18
IDH	Isocitrate dehydrogenase	19
AKGDH	Alpha-ketoglutarate dehydrogenase	20
SCS	Succinyl-CoA synthetase	21
SDH	Succinate dehydrogenase	22
FH	Fumarase	23
MDH	Malate dehydrogenase	24
G6PD	Glucose-6-phosphate dehydrogenase + Gluconolactonase	25
PGD	6-phosphogluconate dehydrogenase	26
RPE	Phosphopentose Epimerase	27
RPI	Phosphopentose Isomerase	28
TKT	Transketolase	29
TALDO	Transaldolase	30
TKT	Transketolase	31

Table E.2 continued.

Abbr.	Enzyme	Rxn
GALK	Galactokinase	32
GALT	Galactose-1-P uridylyltransferase	33
GALE	UDP-Galactose-4-epimerase	34
PGM	Phosphoglucomutase	35
UGP	UDP-Glucose pyrophosphorylase	36
GFPT	Glucosamine-6-P synthase	37
HK	Hexokinase	38
-	Glucosamine N-acetyltransferase	39
NAGK	GlcNAc kinase	40
GNPNAT	Glucosamine-6-P N-acetyltransferase	41
PGM3	Phosphoacetylglucosamine mutase; Phosphoglucomutase 3	42
UAP	UDP-GlcNAc pyrophosphorylase	43
GNE	UDP-GlcNAc-2-epimerase	44
RENBP	N-acylglucosamine 2-epimerase; renin binding protein	45
CTPS	CTP synthetase	46
SIAT	NeuAc transporter	47
-	CTP transporter	48
CMAS	CMP-NeuAc synthetase	49
SLC35A1	CMP-NeuAc transporter	50
CMAH	CMP-NeuAc-4-hydroxylase*	51
PMI	Phosphomannose isomerase	52
PMM	Phosphomannose mutase	53
GMPP	GDP-mannose pyrophosphorylase	54
GMDS	GDP-mannose 4,6-dehydratase	55
Tsta3	GDP-L-fucose synthase; tissue specific transplantation antigen P35B	56
-	GDP-fucose synthetase	57
FPGT	Fucose-1-P guanylyltransferase	58
FUK	Fucokinase	59
GNE	ManNAc kinase	60
NANS	Sialic Acid Synthase	61
NANP	N-acylneuraminate-9-phosphatase	62
-	[GalNAc to GalNAc-1-P]	63
UAP	UDP-GlcNAc pyrophosphorylase	64

Table E.2 continued.

Abbr.	Enzyme	Rxn
-	UDP-GlcNAc-4-epimerase	65
UGDH	UDP-Glc 6-dehydrogenase	66
GLCAK	Glucurokinase	67
-	UDP-GlcA pyrophosphorylase	68
GLUD	Glutamate dehydrogenase	69
-	[a-KG transporter]	70
GALT	Galactose-1-P uridylyltransferase	71
-	N-acetylglucosamine deacetylase	72
Amdhd2	putative N-acetylglucosamine-6-phosphate deacetylase; amidohydrolase domain containing 2	73
CMAH	CMP-N-acetylneuraminate monooxygenase	74
CMAS	N-acylneuraminate cytidylyltransferase	75
HK	hexokinase	76
PAPSS1	3'-phosphoadenosine 5'-phosphosulfate synthase 1	77
PAPSS2	3'-phosphoadenosine 5'-phosphosulfate synthase 2	78
Suox	Sulfite oxidase	79
Ethe	Ethylmalonic encephalopathy 1	80
Cycs	Cytochrome C, somatic	81
Tst	Thiosulfate sulfurtransferase	82
Slc35a1	CMP-NeuAc Transporter	-
Slc35a2	UDP-Gal Transporter	-
Slc35a3	UDP-GlcNAc Transporter	-
Slc35b4	UDP-GlcNAc Transporter	-
Slc35c1	GDP-Fuc Transporter	-
Slc35c2	GDP-Fuc Transporter	-
Slc35d1	UDP-GalNAc/UDP-GlcA Transporter	-
Slc35d2	UDP-Glc/UDP-GlcNAc/GDP-Man(?) Transporter	-
Slc35b2	[PAPS Transporter]	-
Slc35b3	[PAPS Transporter]	-

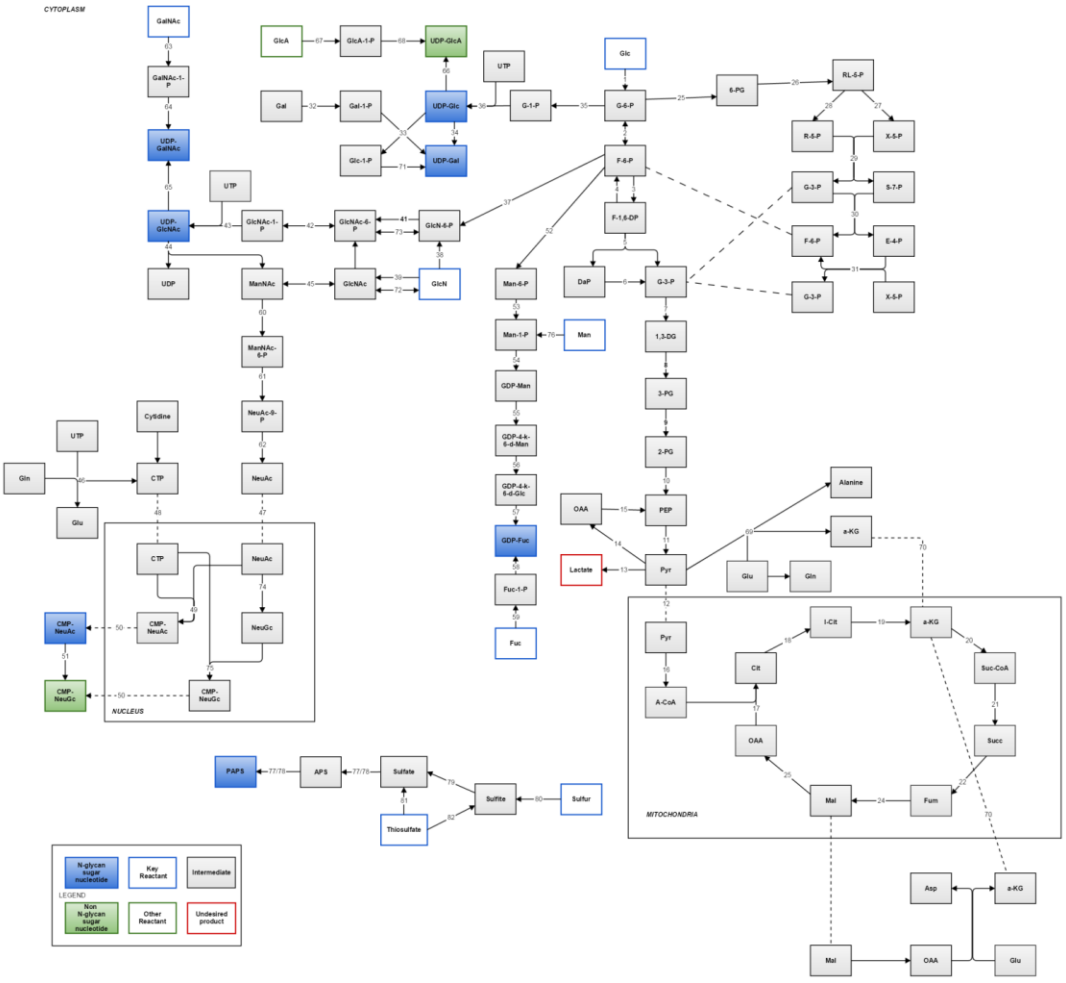


Figure E.1: CHO genome based CCM and sugar nucleotide production pathways.

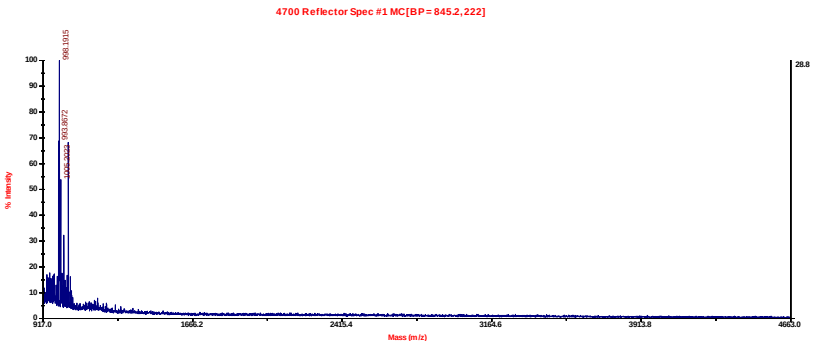
Appendix F

SUPPLEMENTAL EXPERIMENTAL RESULTS FROM CHAPTER 4

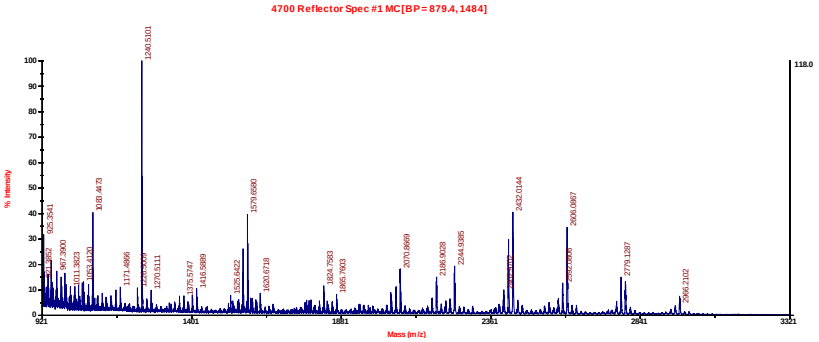
F.1 Preface

As described in Chapter 4, the *GnT-II* knockdown was performed on a CHO-SEAP cell line to confirm a novel Glyco-Mapper prediction. Glycan spectra are given in Figures F.1 and F.2, and qRT-PCR results are given in Figure F.3.

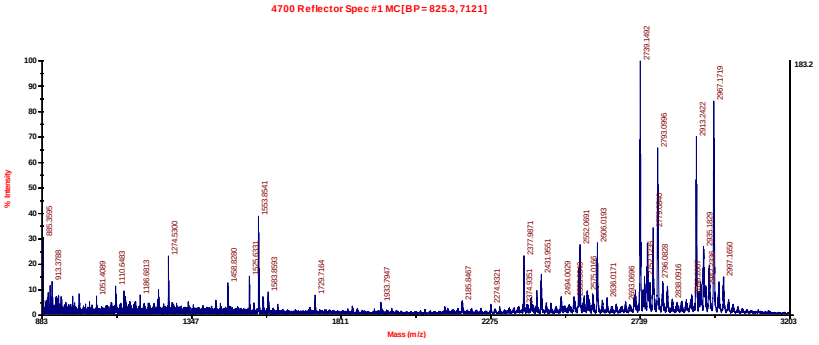
A

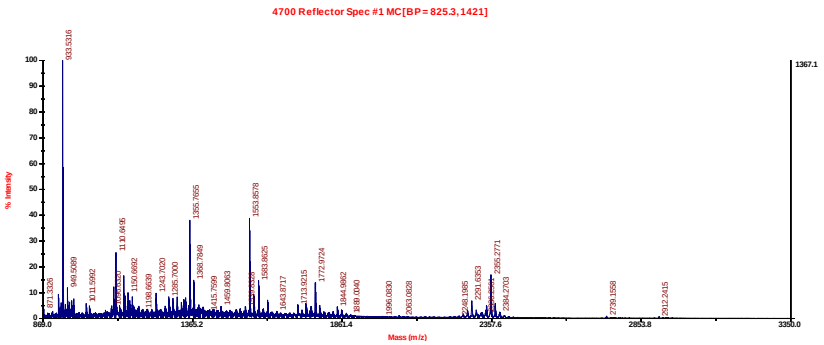

4700 Reflector Spec #1 MC[BP = 845.2, 222]
% Intensity
Mass (m/z)

B


4700 Reflector Spec #1 MC[BP = 879.4, 1484]
% Intensity
Mass (m/z)

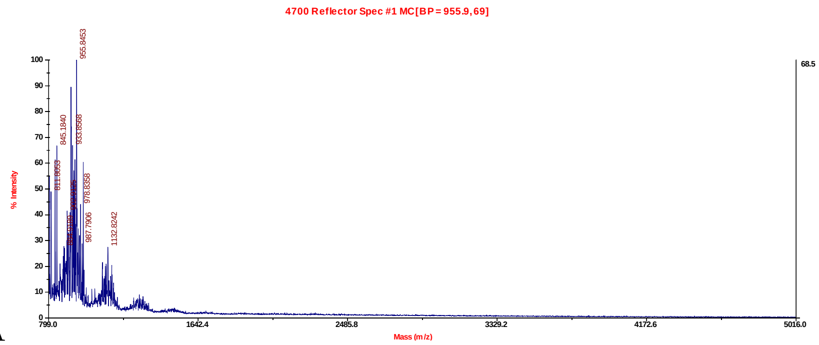
C


4700 Reflector Spec #1 MC[BP = 825.3, 7121]
% Intensity
Mass (m/z)

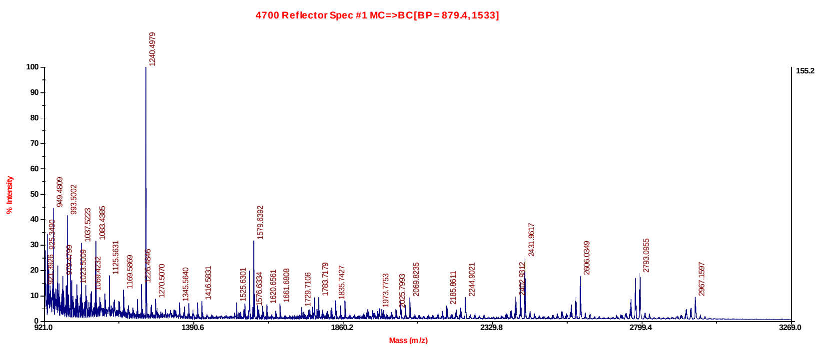


D

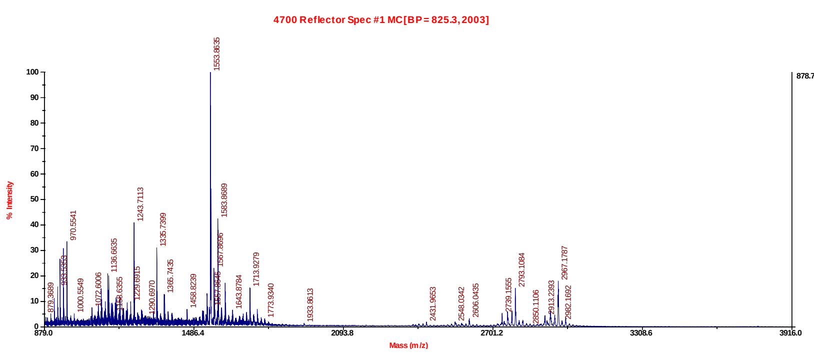
Figure F.1: Reference SEAP glycoform MALDI-TOF MS spectra in increasing acetonitrile aliquot composition order, specifically the A) 15%, B) 35%, C) 50%, and D) 75% aliquots.


4700 Reflector Spec #1 MC[BP = 955.9, 69]
% Intensity
Mass (m/z)
68.5

A


4700 Reflector Spec #1 MC=>BC[BP = 879.4, 1533]
% Intensity
Mass (m/z)
155.2

B


4700 Reflector Spec #1 MC[BP = 825.3, 2003]
% Intensity
Mass (m/z)
878.7

C

D

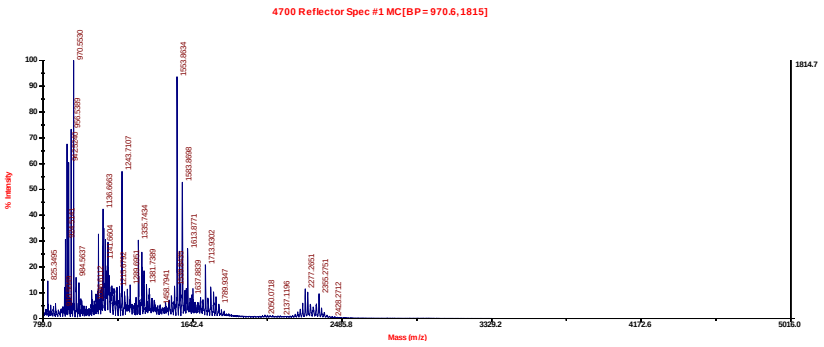


Figure F.2: GnT-II knockdown SEAP glycoform MALDI-TOF MS spectra in increasing acetonitrile aliquot composition order, specifically the A) 15%, B) 35%, C) 50%, and D) 75% aliquots.

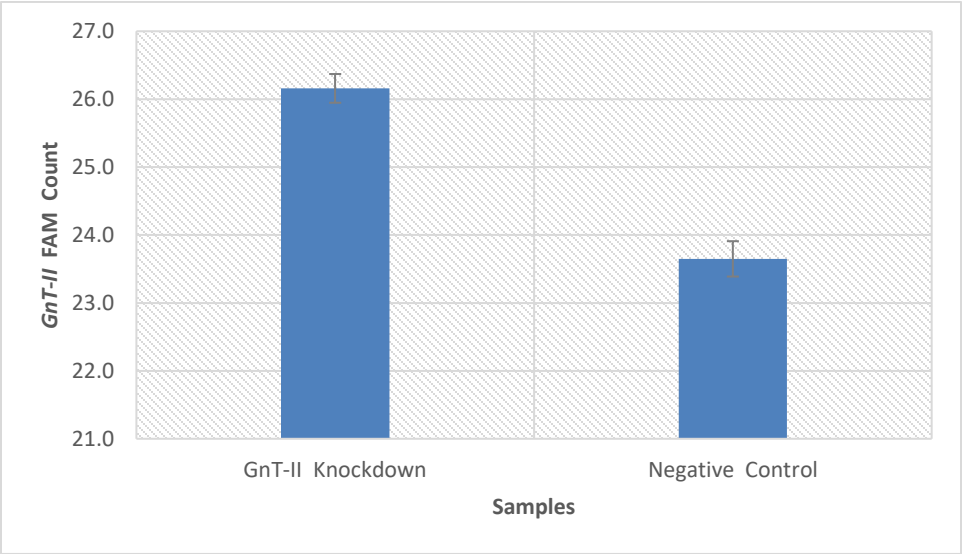


Figure F.3: qRT-PCR results demonstrating a statistically significant knockdown ($p<0.0001$) of *GnT-II* via siRNA by an average increase of 2.51 compared to the negative control. n=9 for each sample with 3 technical replicates of 3 biological replicates.